

Comparative genomics tools for biological discovery

Inna Dubchak, Ph.D.

Berkeley PGA, Bioinformatics Group Leader

Lawrence Berkeley National Laboratory

ildubchak@lbl.gov

<http://www-gsd.lbl.gov>

Outline

What is comparative genomics?

VISTA tools developed for comparative genomics.

Related biological stories

Large scale VISTA applications including automatic computational system for comparing whole vertebrate genomes

Human genome 2001



Fugu genome 2002

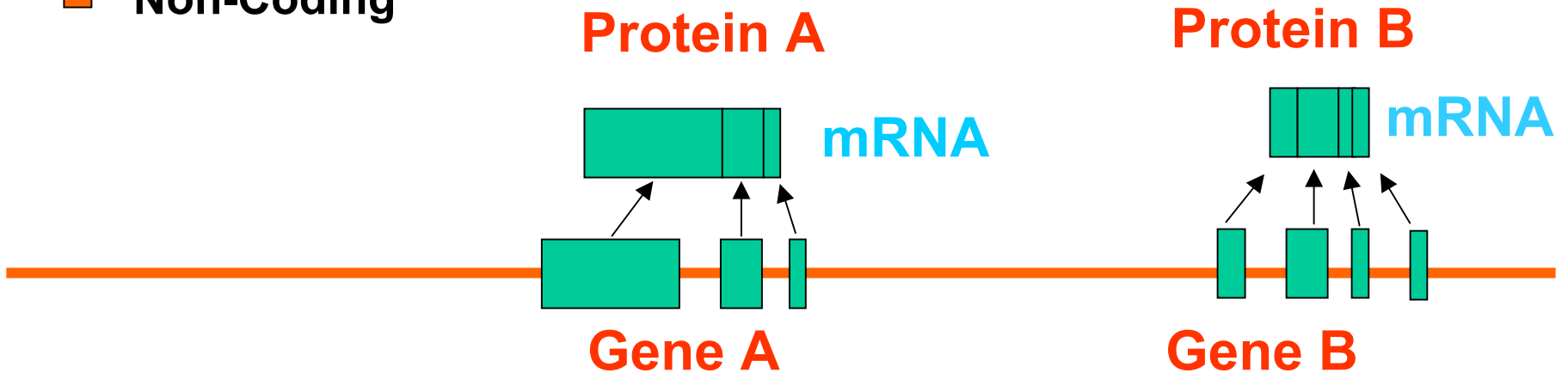


Mouse genome 2002



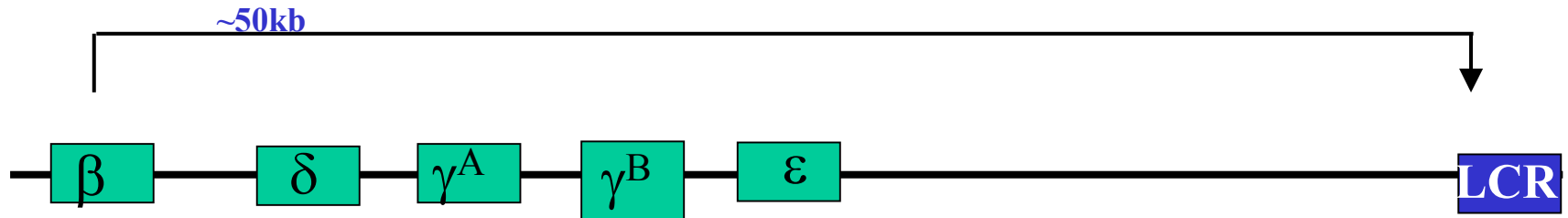
1-2% Coding

-  Coding
-  Non-Coding



Distant Non-Coding Sequences Causing Disease

β -Thalassemia

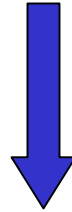


Disease	Gene	Distance
Campomelic dysplasia	SOX9	850kb
Aniridia	PAX6	125kb
X-Linked Deafness	POU3F4	900kb
Saethre-Chotzen syndrome	TWIST	250kb
Rieger syndrome	PITX2	90kb
Split hand/split foot malformation	SHFM1	450kb

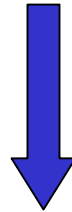
Background

Evolution can help!

In general, functionally important sequences are conserved



Conserved sequences are functionally important



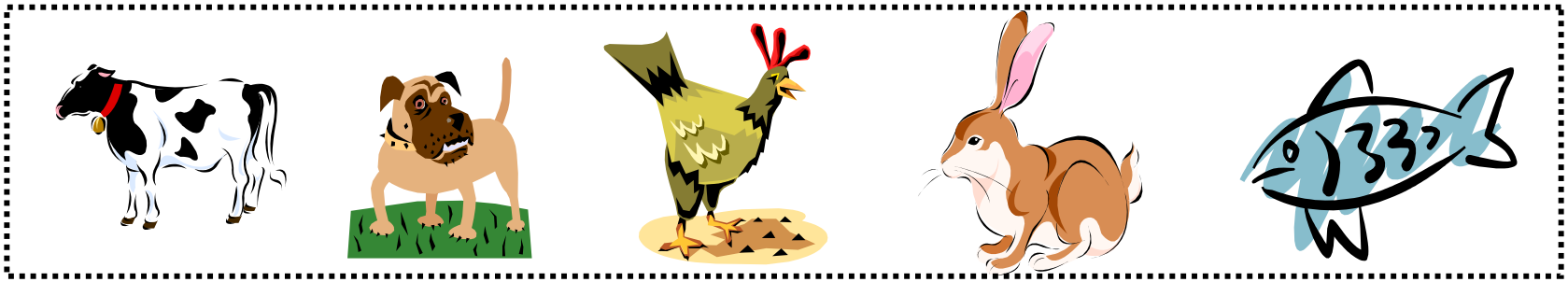
Raw sequence can help in finding biological function

Comparison of 1196 orthologous genes (Makalowski et al., 1996)

- Sequence identity:
 - exons: 84.6%
 - protein: 85.4%
 - introns: 35%
 - 5' UTRs: 67%
 - 3' UTRs: 69%
- 27 proteins were 100% identical

Integrating data into more powerful gene prediction models than with human genomic sequence alone

Comparing sequences of different organisms



- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

Sequence comparisons. How?

Three variations:

Find the best **OVERALL** alignment.

Global alignment

Find **ALL** regions of similarity.

Local alignment

Find the **BEST** region of similarity.

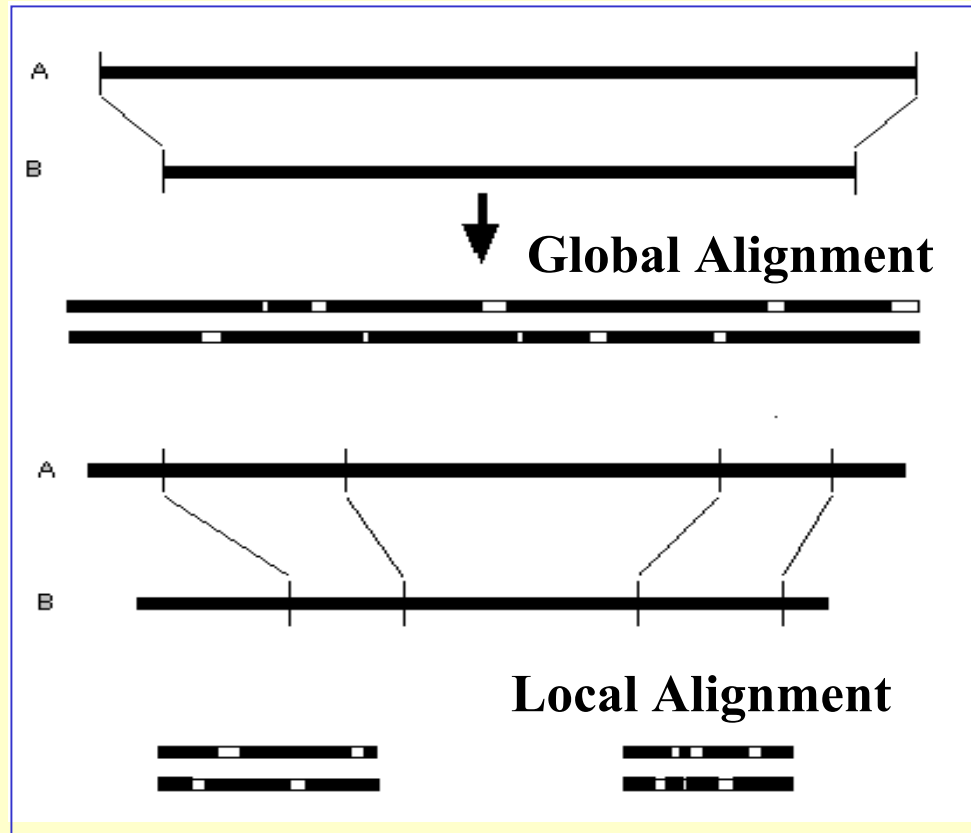
Optimal local alignment

Local alignment algorithms are designed to search for highly similar regions in two sequences that may not be highly similar in their entirety. The algorithm works by first finding very short common segments between the input sequence and database sequences, and then expanding out the matching regions as far as possible.

For cross-species comparison one needs to accurately align two complete sequences. It is insufficient to find common similar regions in the two sequences, rather, what is needed is a global map specifying how the two sequences fit together, much like understanding how the pieces in a puzzle connect up with each other.

This problem is called global alignment

Local vs global alignment



Challenges in aligning long genomic regions

- Long sequences lead to memory problems
- Speed becomes an issue
- Long alignments are very sensitive to parameters
- Draft sequences present a nontrivial problem
- Accuracy is difficult to measure and to achieve
- Scaling up to the size of whole genomes
- Sequence at different stages of completion, difficult to compare

Whole genome shotgun  Partial Assemblies
Finished BACs

<http://www-gsd.lbl.gov/vista>



VISUALIZATION **T**OOLS FOR **A**LIGNMENTS

VISTA

USE VISTA on the WEB

Vista -- [instructions](#) for using **VISTA**

rVista -- [instructions](#) for using **rVISTA**

DOWNLOAD VISTA

Vista Go to our [software download page](#) to obtain **VISTA**'s alignment and visualization programs.

INFORMATION about VISTA

Vista How to [cite VISTA](#).
[Send us your questions, comments](#)

WELCOME to the homepage for **VISTA**, **V**isualization **T**ool for **A**lignments.

Vista is an integrated computational system for global alignment and visualization, designed for comparative genomics. It allows for the visualization of long sequence alignments of DNA from two or more species with annotation information, and it was developed to locate conserved sequences in syntenic regions ([Dubchak et al., 2000](#)).

It has a clean output, allowing for easy identification of sequence similarities and differences, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution.

This system consists of several unified modules:

avld the program for global alignment of DNA sequences of arbitrary length. In addition to aligning two finished sequences, it can also handle one sequence in a non-ordered and non-oriented draft format [Details](#).

Vista A computational tool for comparing an arbitrary number of genomic sequences from different species. [Details](#)

Modules of VISTA:

- Program for global alignment of DNA fragments of any length (AVID)
- Visualization of alignment and various sequence features for any number of species
- Evaluation and retrieval of all regions with predefined levels of conservation

Visualization



tggtaacattcaaattatg-----ttctcaaagtgagcatgaca-acttttttccatgg
|| | |||| | | || || | | | | | | | | | | | | | | | |
tgatgacatctatttgctgtttccttttttagaaactgcatgagagcctggctagtaggg



Window of length **L** is centered at a particular nucleotide in the base sequence

Percent of identical nucleotides in **L** positions of the alignment is calculated and plotted

Move to the next nucleotide

Finding conserved regions with percentage and length cutoffs

Conserved segments with percent identity X and length Y - regions in which every contiguous subsegment of length Y was at least $X\%$ identical to its paired sequence. These segments are merged to define the conserved regions.

Output:

11054 - 11156 = 103bp at 77.670%	NONCODING
13241 - 13453 = 213bp at 87.793%	EXON
14698 - 14822 = 125bp at 84.800%	EXON

VISTA input files

Sequences

```
> Human ST7 gene
CTGAATGGCTCGTAGAAA
TATTGCATTAACCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. .
```

```
> Cow ST7 gene
CTGAATGGCTCGTAGAAA
TAATGCATTCCCCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. . . .
```

```
. . . . .
```

Annotation for a base sequence if available

```
> 12877 289557 ST7b/a
+ 13076 282515
12877 13226
159297 159379
179096 179255
189328 189382
```

VISTA output files

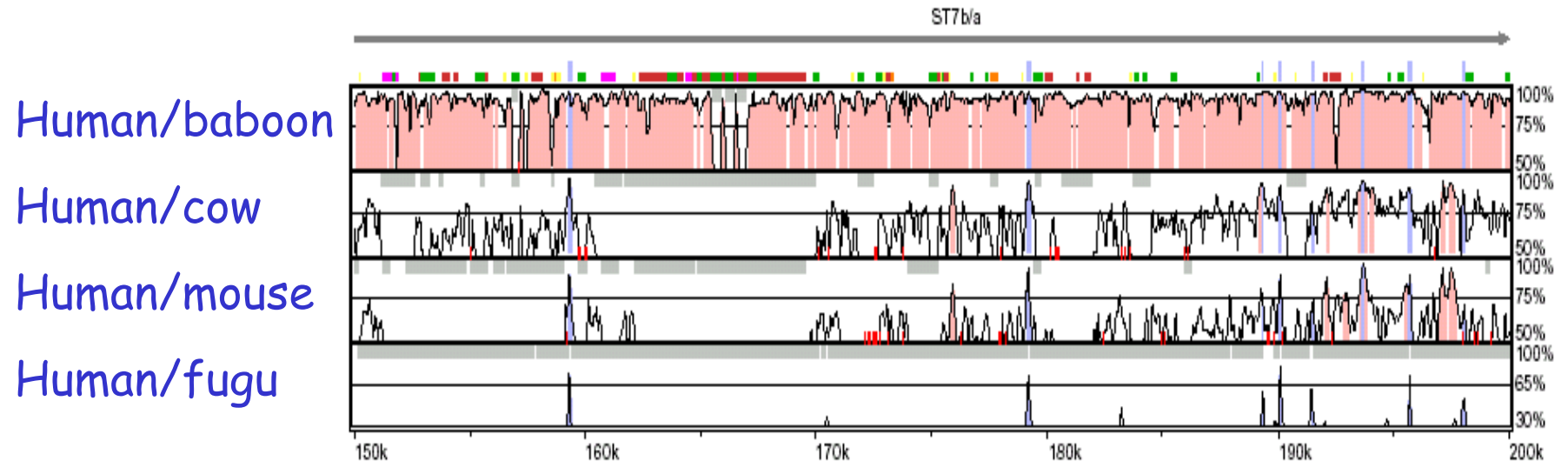
All pair wise alignments

```
185140    185150    185160                185170    185180
GACATTGGAAAAGTAAAGGAAGTGGTTTAT---CTTGCTC-----TTTTTGCAACAGTA
      ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
GACACTGGAAAAGCAGAGGAAGTGGTTTATTGACCTGCCCCCCCCTTTTTTATAACAGTG
```

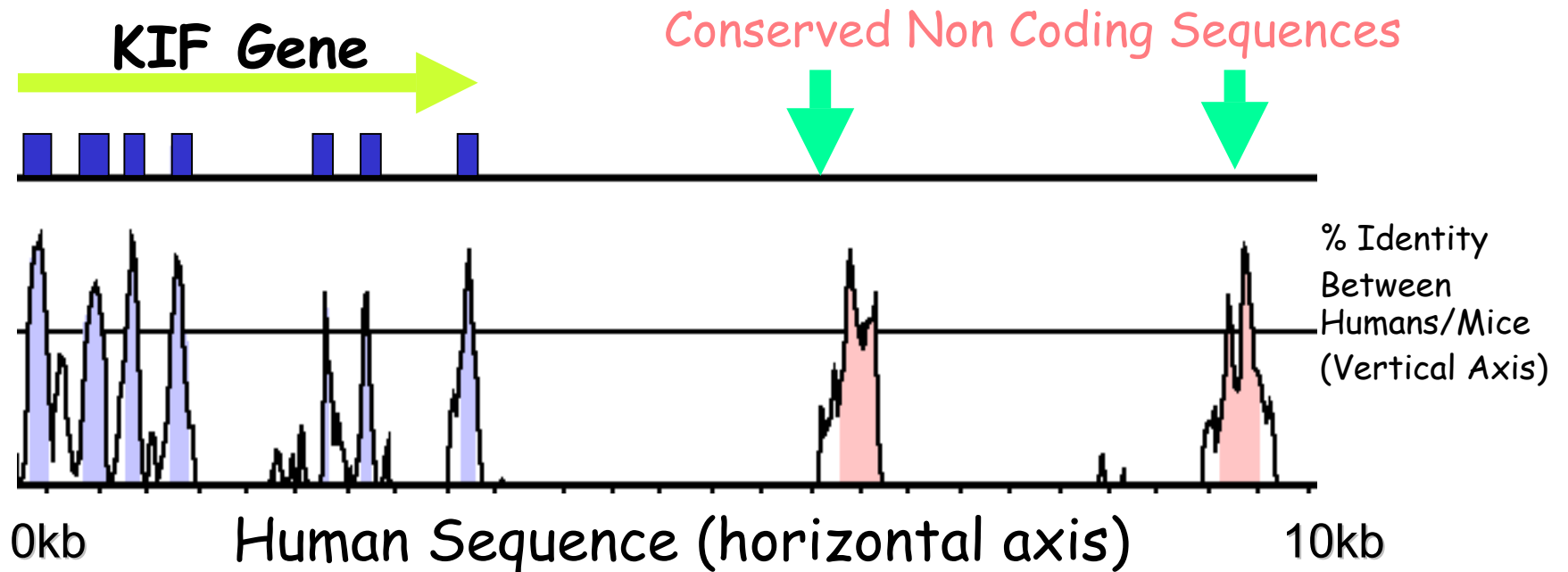
The lists of conserved regions

```
80078 (149626) to 80171 (149724) = 99bp at 63.6% noncoding
159297 (158141) to 159379 (158223) = 83bp at 80.7% exon
179096 (159067) to 179253 (159224) = 158bp at 75.9% exon
189328 (159566) to 189382 (159620) = 55bp at 81.8% exon
```

VISTA plot



VISTA plot



<http://www-gsd.lbl.gov/vista>



> 30000 queries on-line, distributed > 1250 copies of the program in 48 countries.

After VISTA publications at the end of 2000:

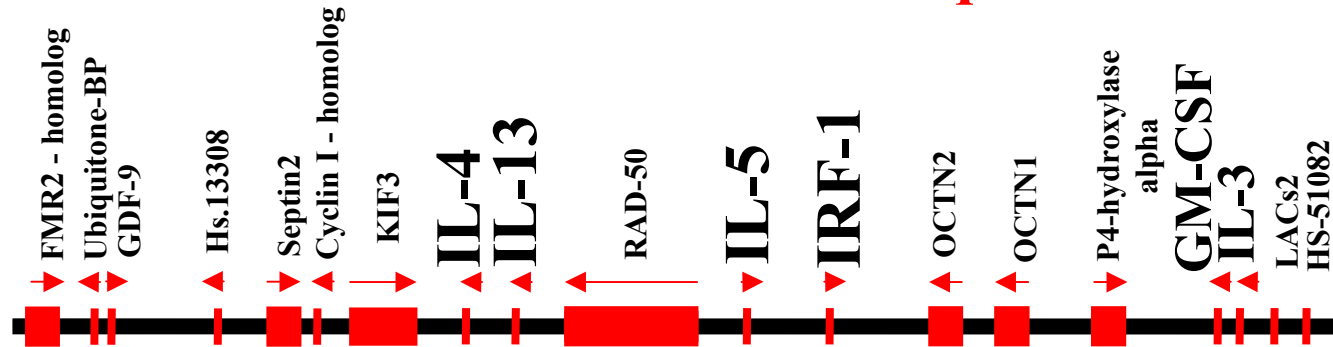
~60 papers cited VISTA and presented results obtained with the program

Biological story

Discovering Interleukin Expression Switch

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*. 2000 Apr 7;288(5463):136-40.

IL Cluster HUM 5q31

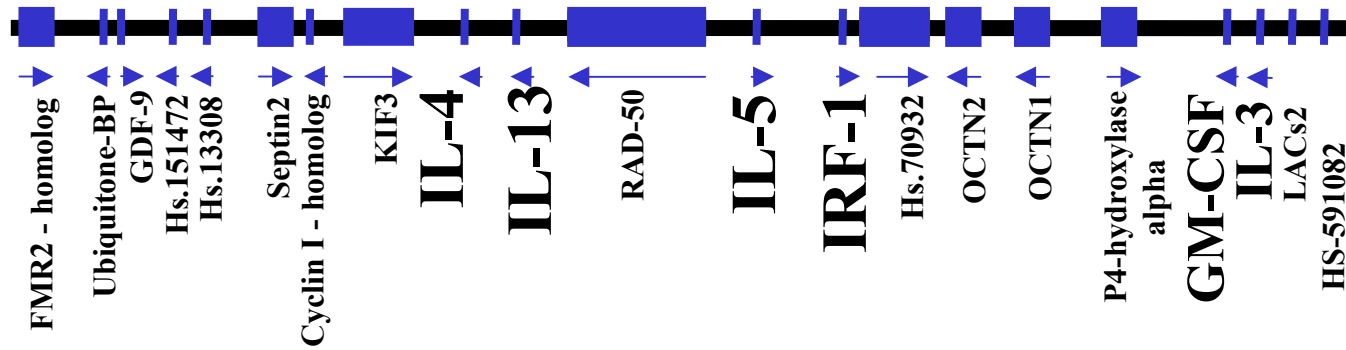


Coding

Exons 3%

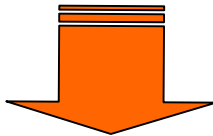
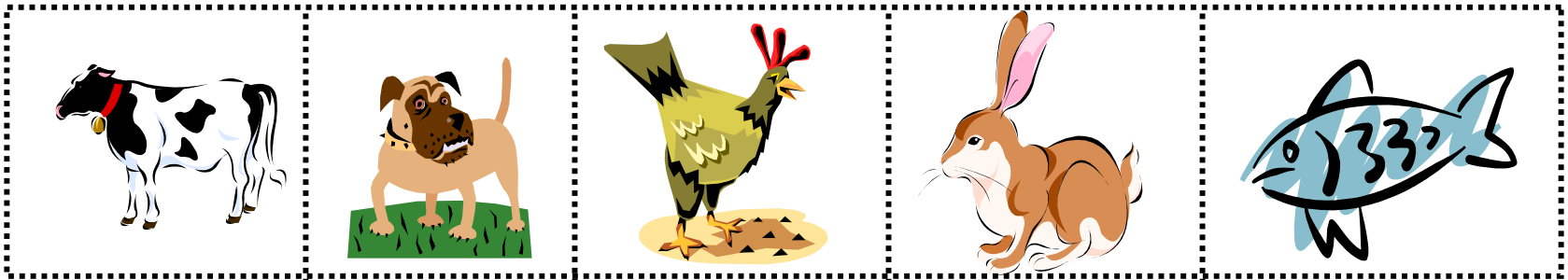
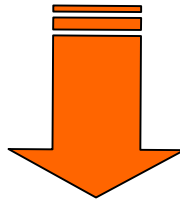
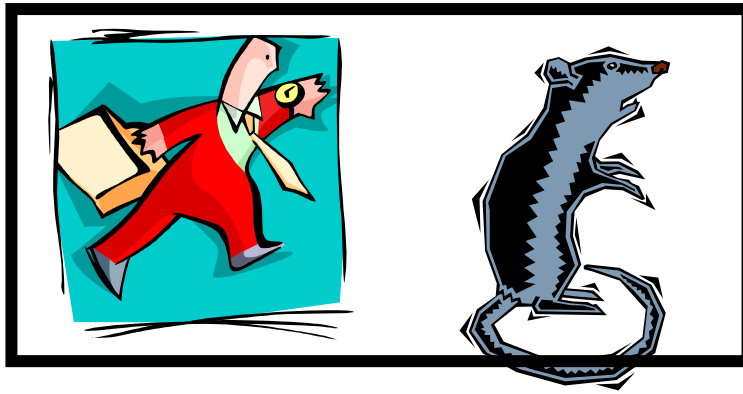
Non-Coding

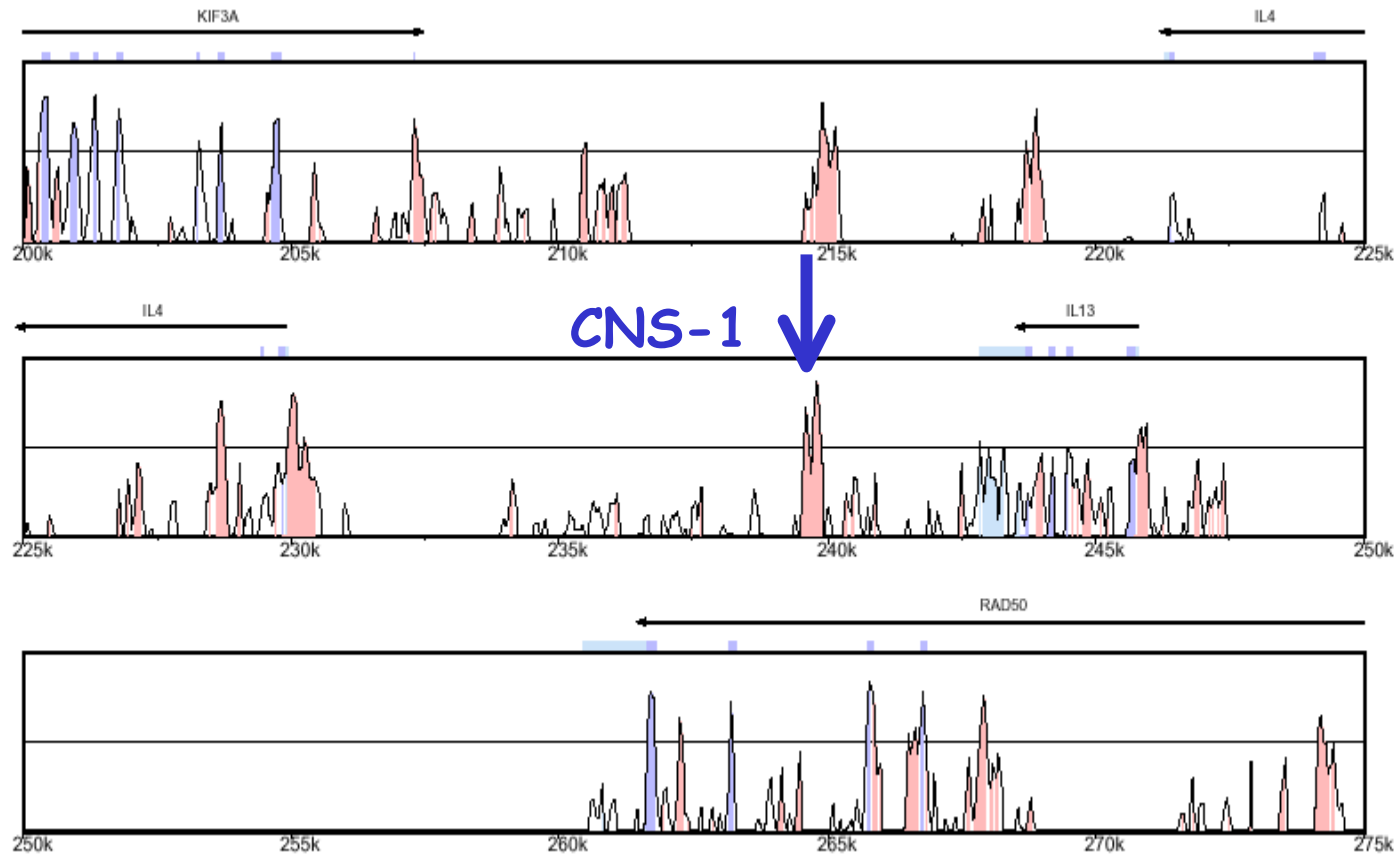
Conserved 2.6%
(>100bp > 75%)



IL Cluster MU Ch 11

A Filtering Strategy





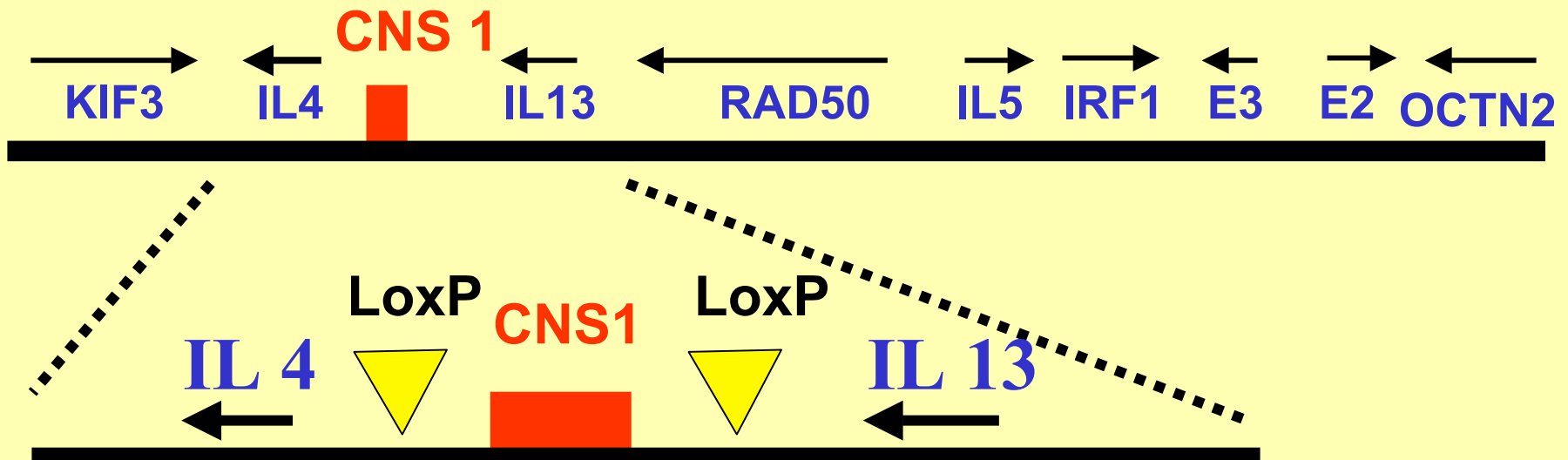
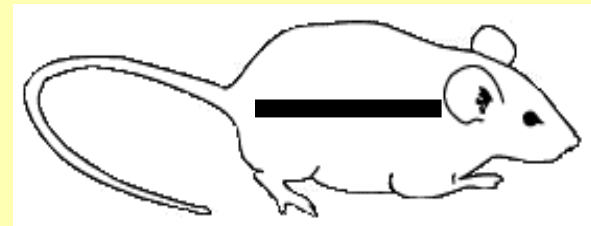
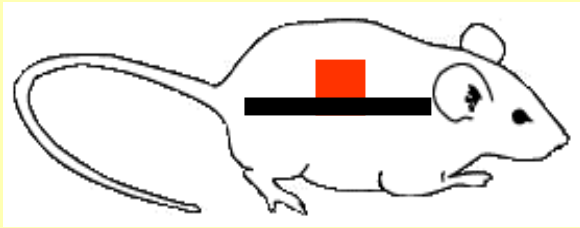
Present in other species: Cow (86%), Dog (81%), Rabbit (73%)

Genomic position conserved in human, mouse, dog, baboon

Single copy in the human genome. Two hypersensitive sites mapped.

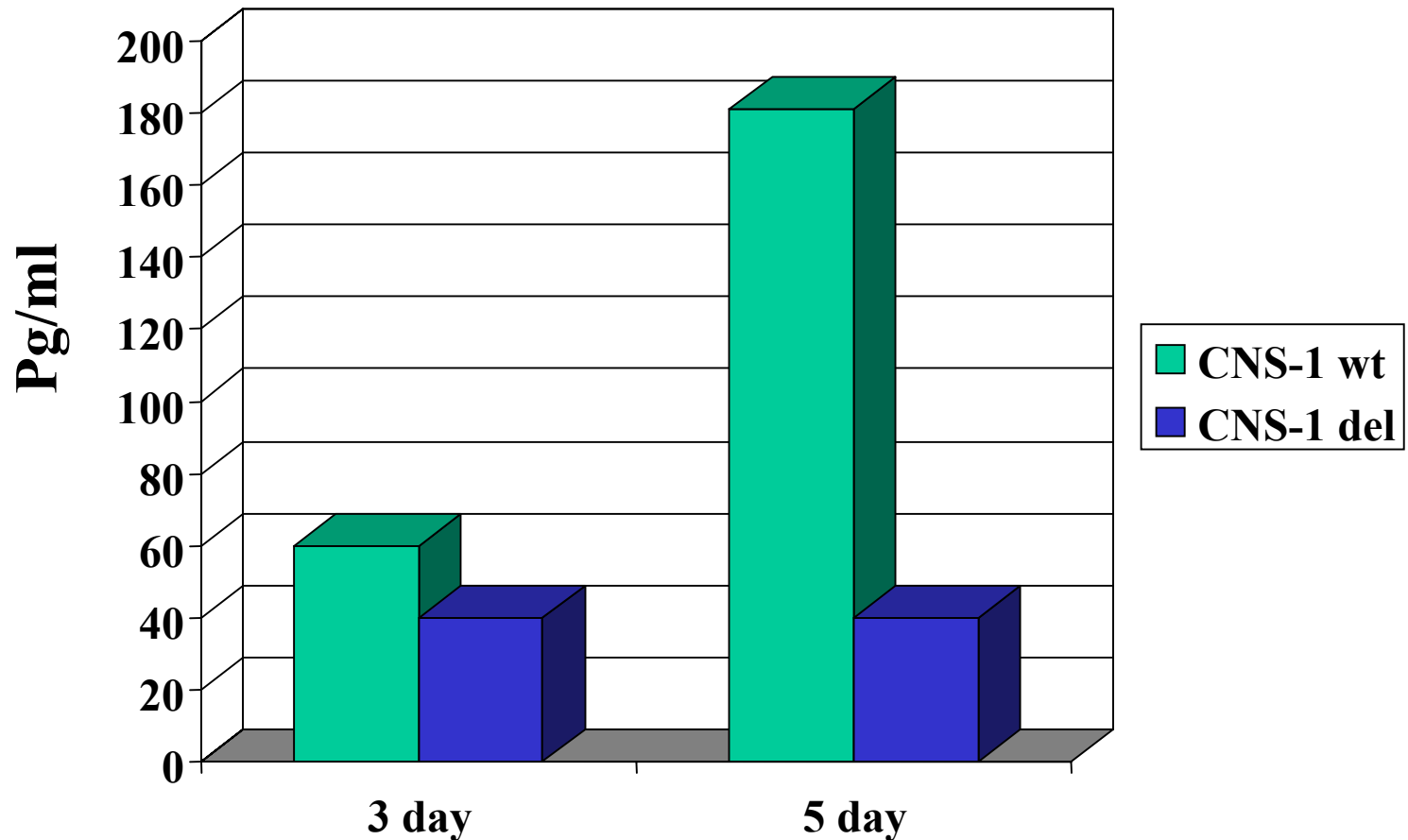
Functional Analysis of CNS1

Generate Human 5q31 YAC Transgenic Mice



Human IL 4 Production in YAC Transgenics Containing and Lacking CNS1

IL-5 & IL13 Expression is also reduced in CNS-1^{del} mice



KIF3

IL13

IL5

IL4

CNS-1

RAD50



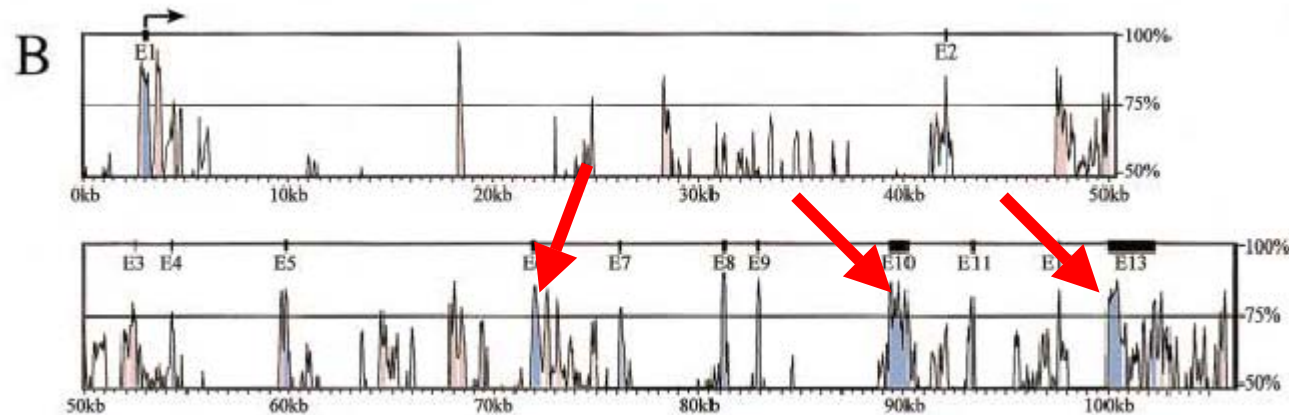
Results obtained with VISTA

J Mol Cell Cardiol 34, 1345-1356 (2002)

Myocardin: A Component of a Molecular Switch for Smooth Muscle Differentiation. J. Chen, C. M. Kitchen, J. W. Streb and J. M. Miano

University of Oxford

VISTA used to solve the **gene structures** of rat and human myocardin.

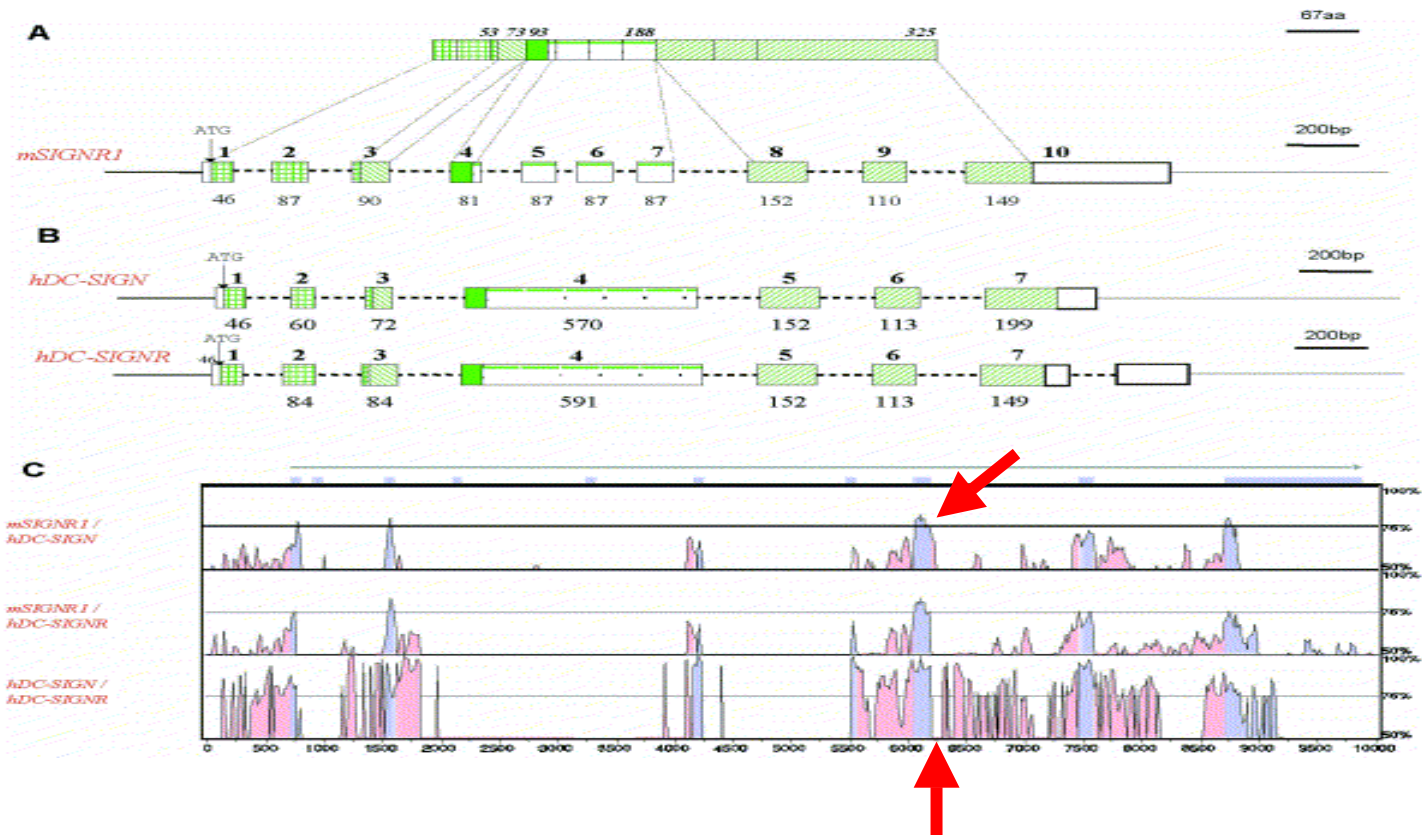


Gene 293, 33-46 (2002)

Molecular characterization of the murine SIGNR1 gene encoding a C-type lectin homologous to human DC-SIGN and DC-SIGNR

S. A. Parent, T. Zhang, G. Chrebet, J. A. Clemas, D. J. Figueroa, B. Ky, R. A. Blevins, C. P. Austin and H. Rosen

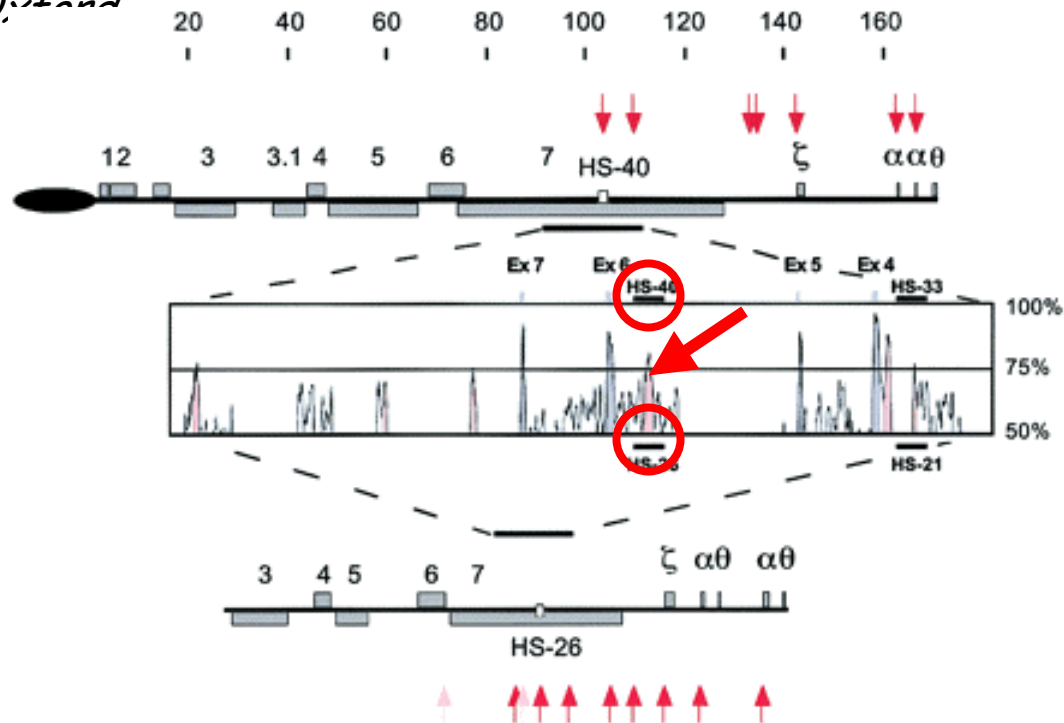
Merck Research Laboratories.



Blood, 100, 3450-3456 (2002)

Deletion of the mouse α -globin regulatory element (HS 26) has an unexpectedly mild phenotype

E. Anguita, J. A. Sharpe, J. A. Sloane-Stanley, C. Tufarelli, D. R. Higgs, and W. G. Wood
University of Oxford



(HS 40) is necessary for high-level expression of the α -globin genes. A similar element in the mouse (mHS 26) supposedly has similar functional properties. Knock out mHS26 instead of the expected severe α -thalassemia phenotype, produce the mice with a mild disease. These results may indicate differences in the regulation of the α -globin clusters in mice and humans.

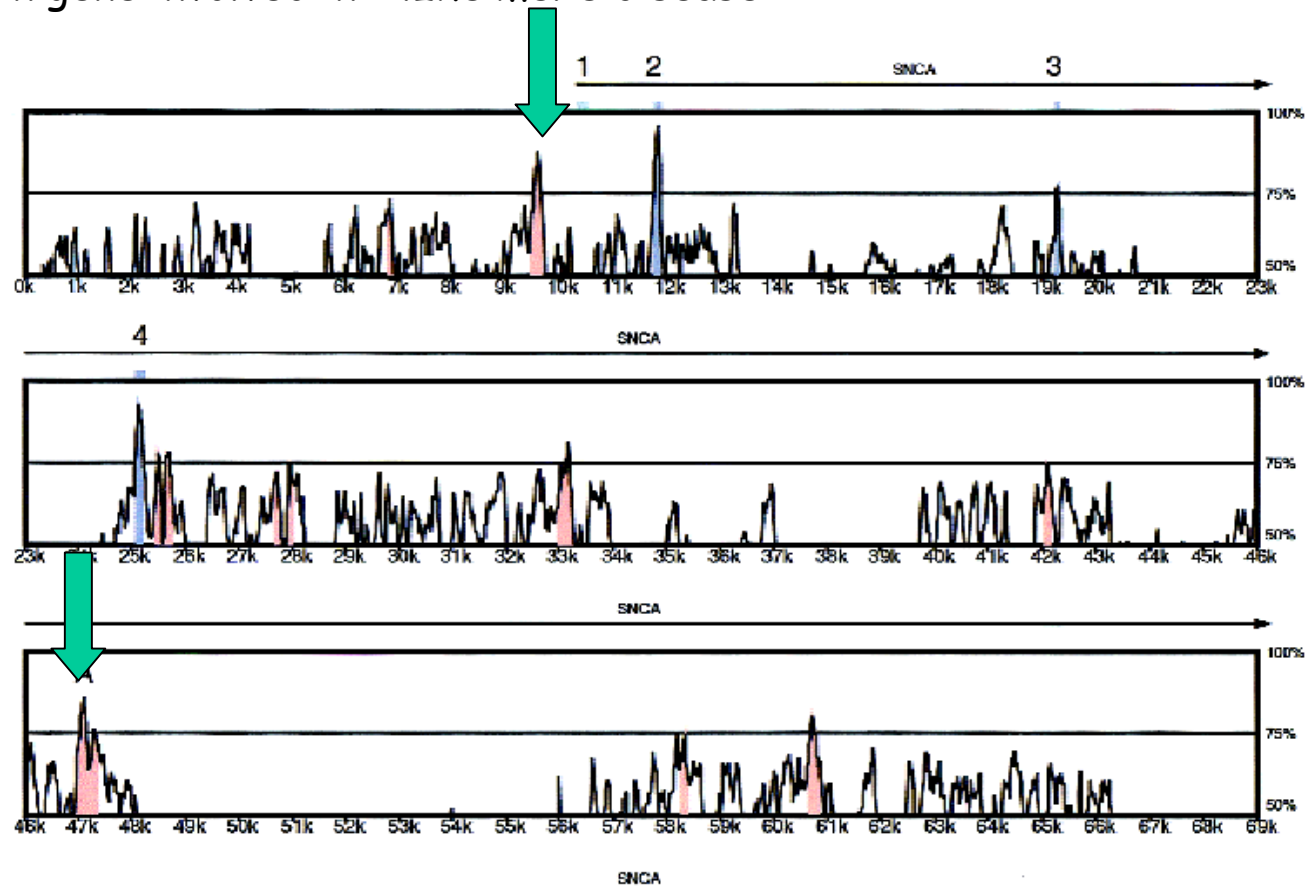
Genome Research 11, 78 (2001)

Human and Mouse - Synuclein Genes: Comparative Genomic Sequence Analysis and Identification of a Novel Gene Regulatory Element

J. W. Touchman, et al.

NIH Intramural Sequencing Center, National Institutes of Health

Synuclein gene involved in Alzheimer's disease



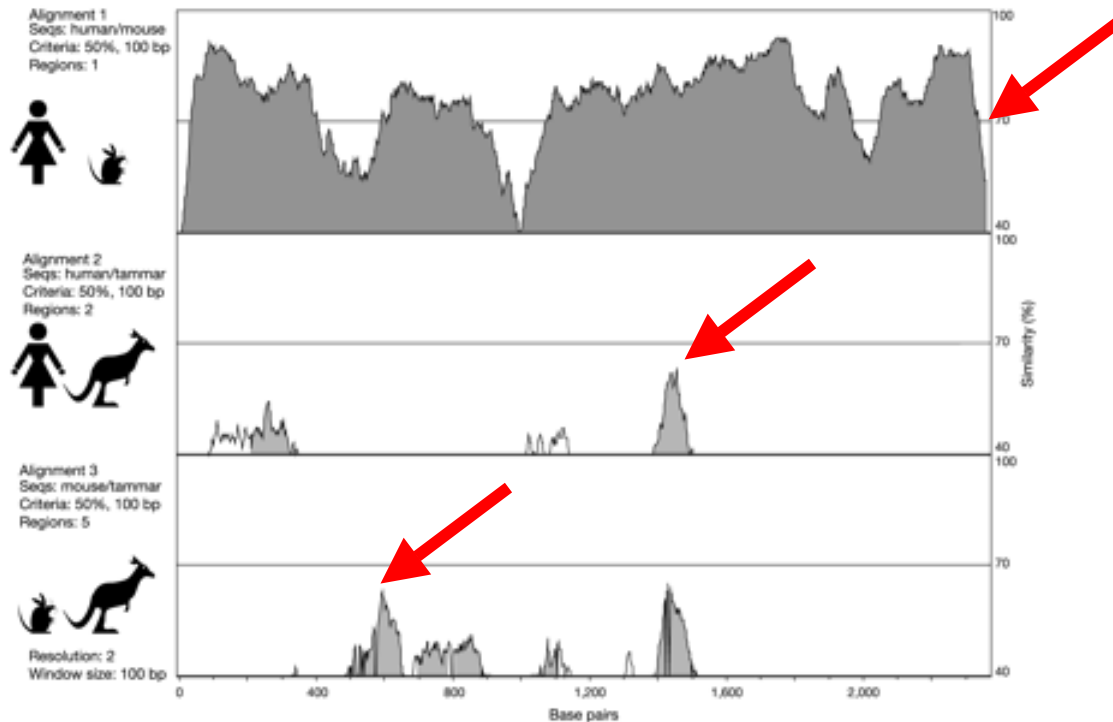
EMBO reports 4:143 (2003)

The kangaroo genome. Leaps and bounds in comparative genomics

M. J. Wakefield and J. A. Marshall Graves

Research School of Biological Sciences, The Australian National University,
Canberra, ACT 0200, Australia

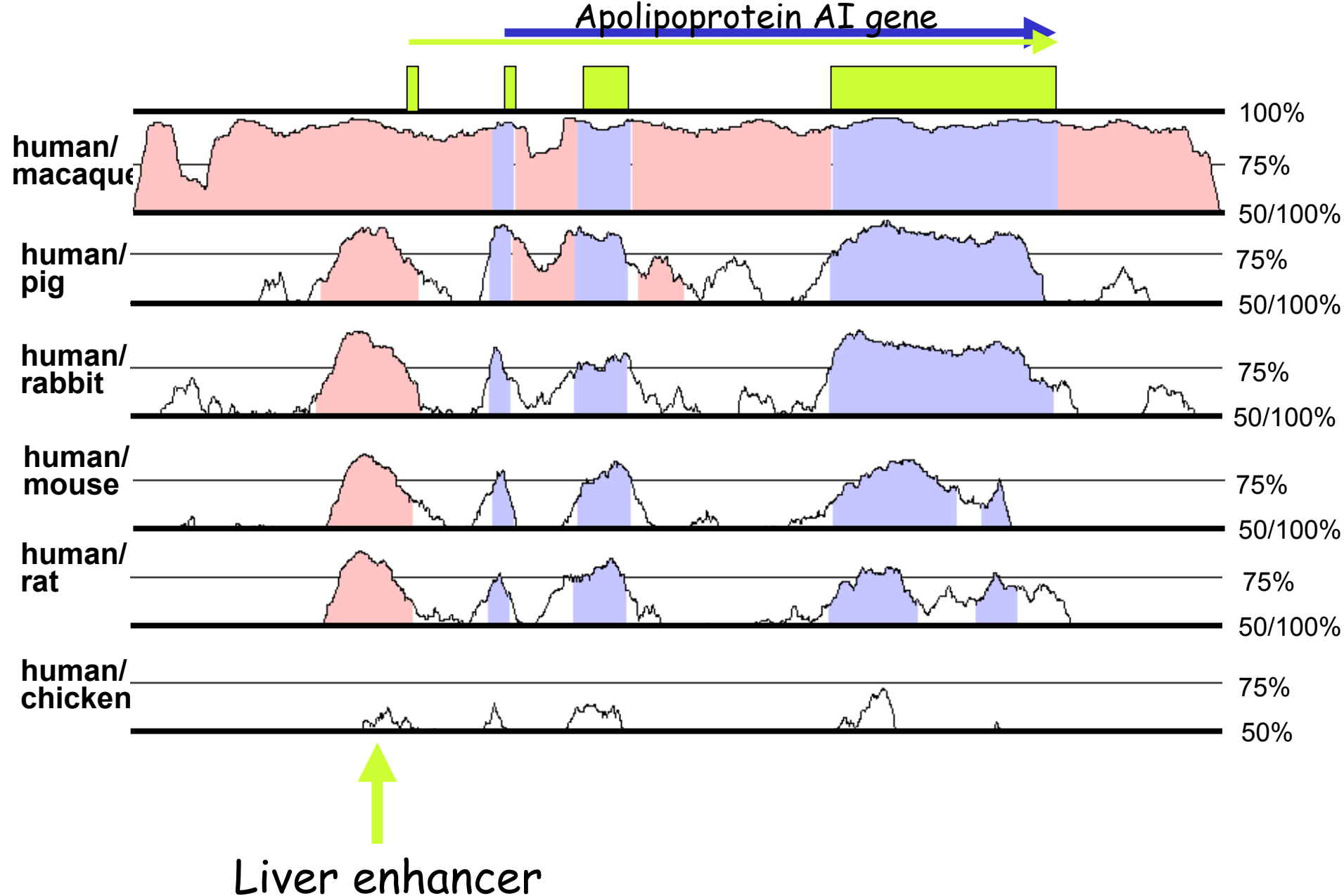
'The kangaroo genome is a rich and unique resource for comparative genomics,
a treasure trove of comparative genomics data'.



Phylogenetic footprinting of 3' untranslated region of the SLC16A2 gene

Multi-Species Comparative Analysis (VISTA)

Apolipoprotein AI gene



VISTA family of tools

<http://www-gsd.lbl.gov/vista>

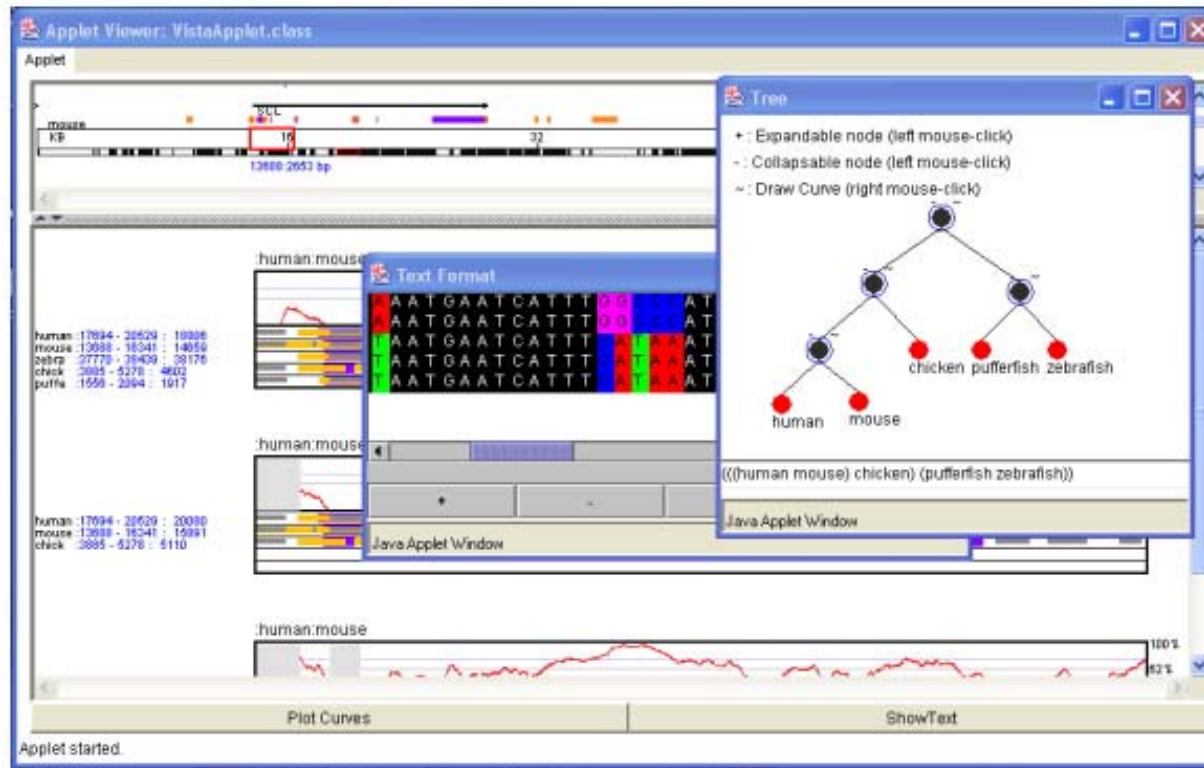
- **VISTA** - comparing DNA of multiple organisms
- **for 3 species** - analyzing cutoffs to define actively conserved non-coding sequences
- **cVISTA** - comparing two closely related species
- **PhyloVISTA** - visualization of multiple sequence alignments in the context of their evolutionary relationship
- **rVISTA** - regulatory VISTA

PhyloVISTA

Multiple alignments Visualization

- more and more genomes sequenced
- multiple alignments reveals conservations, mutation, deletions, duplications events across the phylogenetic tree
- need comparison in the context of the phylogenetic relationship
- need to understand conservation down to motifs

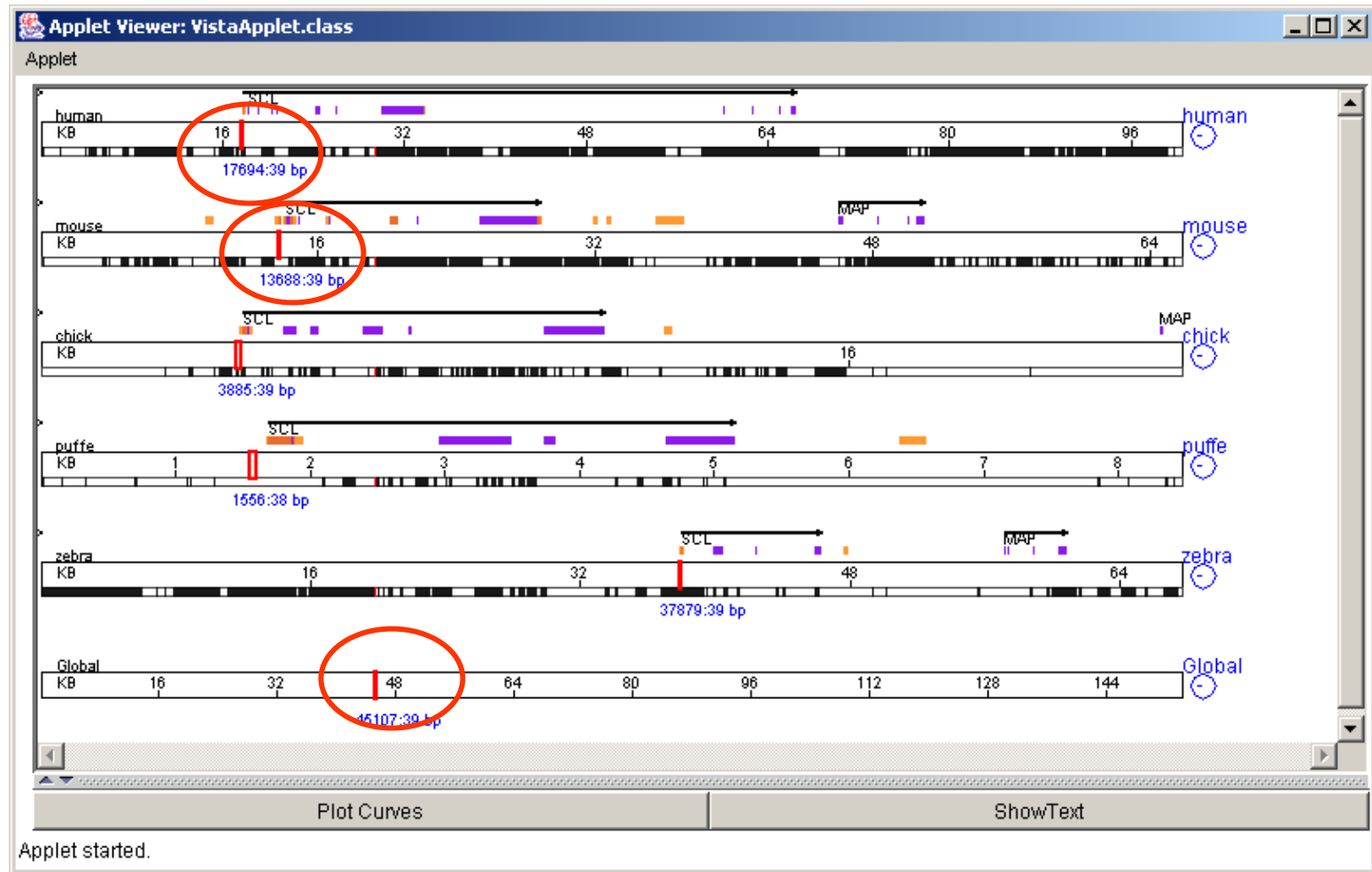
PhyloVISTA



www-gsd.lbl.gov/phylovista

Shah et al (2003) Bioinformatics, submitted

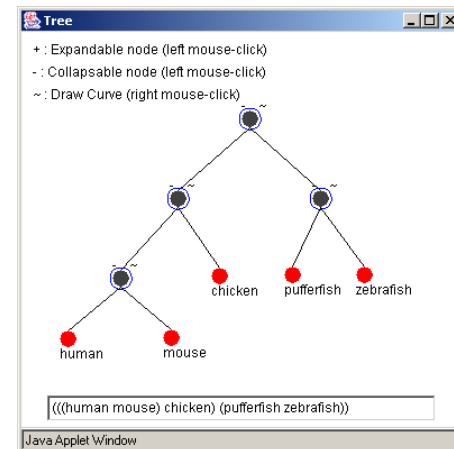
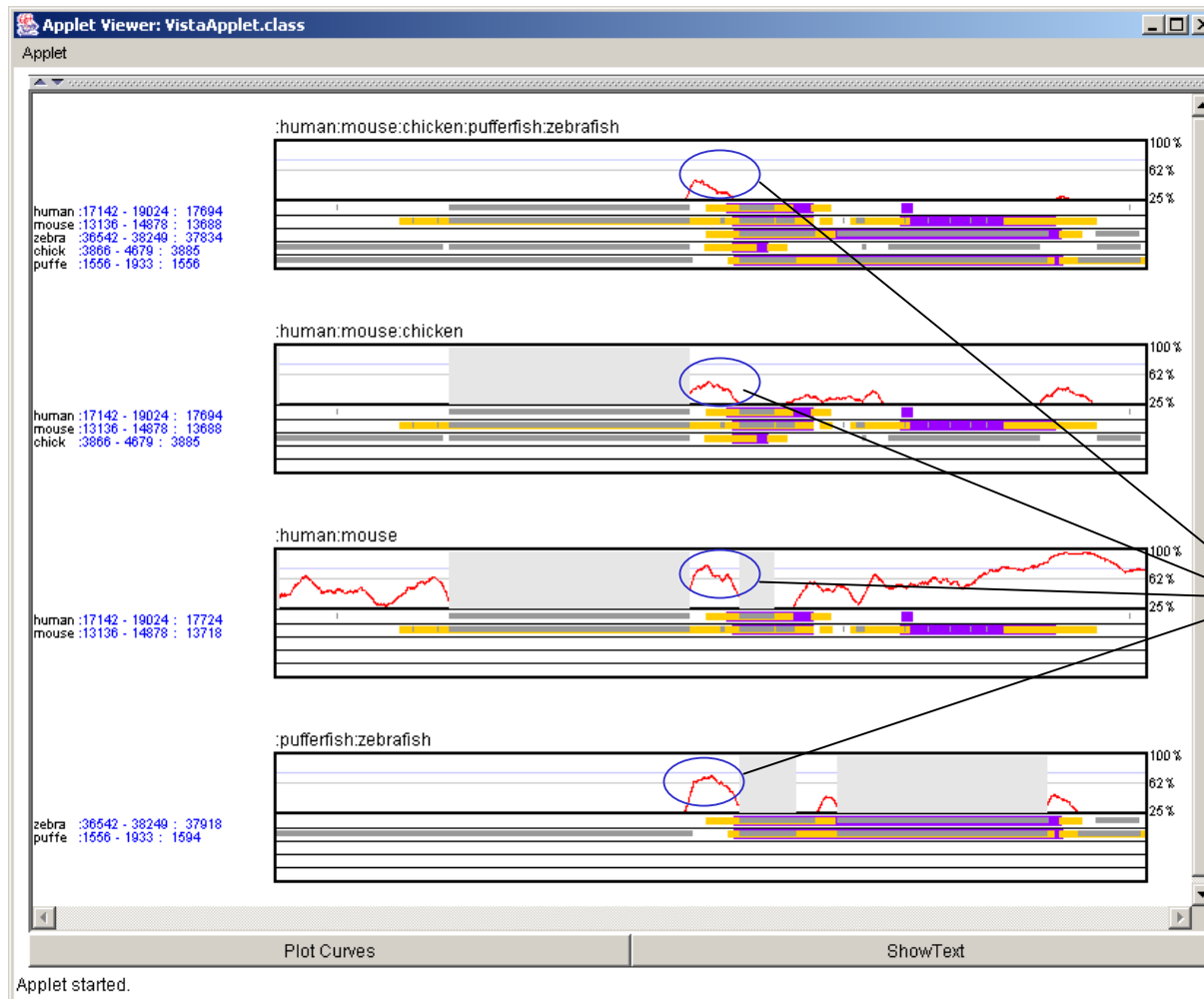
Visualization of Multiple alignment - Phylo VISTA



<http://www-gsd.lbl.gov/phylovista>

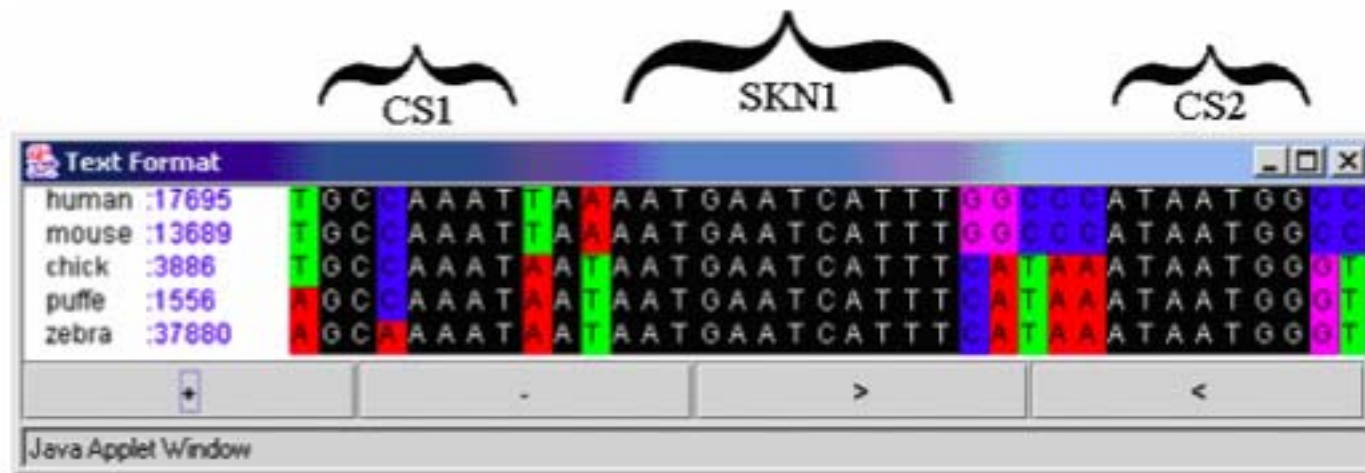
Visualization of Multiple Alignment - Phylo VISTA

Plots of similarity measure for four selected nodes

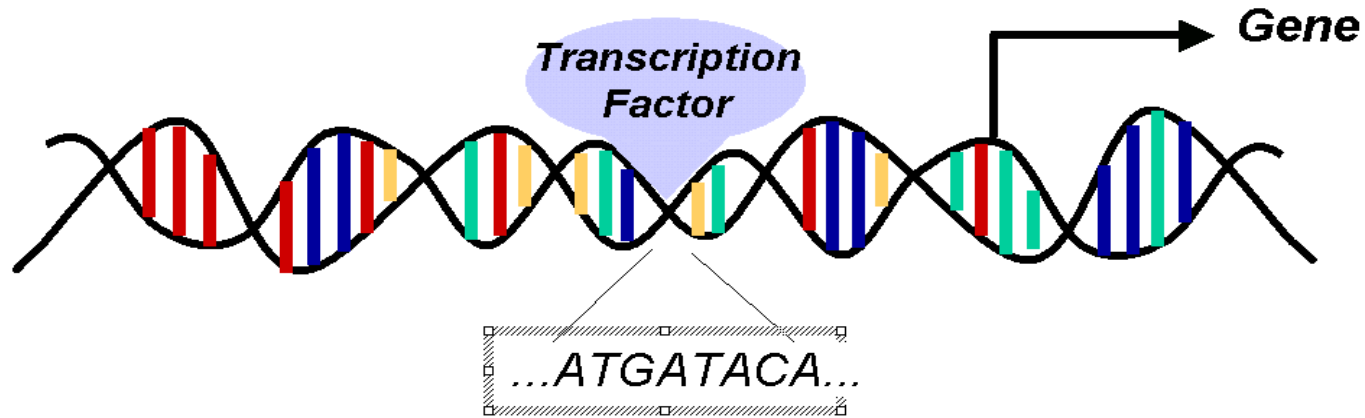


**Peaks indicate
strongly
conserved region.**

visualization of multiple alignments for motif discovery



rVISTA - prediction of transcription factor binding sites



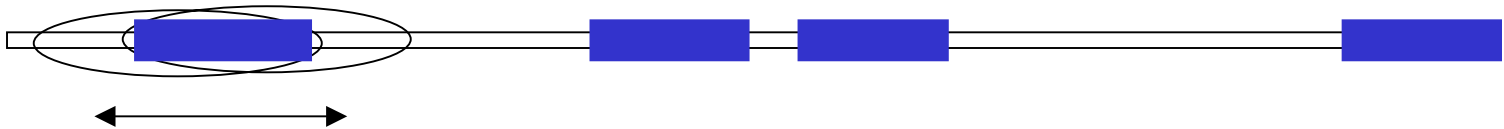
- Simultaneous searches of the major transcription factor binding site database ([Transfac](#)) and the use of global sequence alignment to sieve through the data

Regulatory VISTA (rVISTA)

1. Identify potential transcription factor binding sites for each sequence using library of matrices (TRANSFAC)
2. Identify aligned sites using VISTA
3. Identify conserved sites using dynamic shifting window

Percentage of conserved sites of the total 3-5%

	Ikaros-2	Ikaros-2	NFAT	Ikaros-2
Human	TGATTTCTCG GCAGCAAGGGAGGG CCCCATGACAAAGCCATTT GAAATCCCAGAA GC AATTTTCTACTT ACGACCTCACTTTCTGTTGCTGTCTCT CCCTTCCCCTCTG			
Mouse	TGATTTCTCG GCAGCCAGGGAGGG CCCCATGACGAAGCCACTC GAAATCCCAGAA GCA AATTTTCTACTT ACGACCTCACTTTCTGTTGCTCTCTCT TCCTCCCCCTCCA			
Dog	TGATTTCTCG GCAGCAAGGGAGGG CCCCATGACGAAGCCATTT GAAATCCCAGAA GCA AATTTTCTACTT ACGACCTCACTTTCTGTTGCGTCACT CCCTTCCCCTGCA			
Rat	TGATTTCTCG GCAGCCAGGGAGGG CCCCATGACGAAGCCACTC GAAATCCCAGAA GCA AATTTTCTACTT ACGACCTCACTTTCTGTTGTTCTCTCT TCTCTCCCCCTCCA			
Cow	TGATTTCTCG GCAGCCAGGGAGGG CCCCATGACGAAGCCATTT GAAATCCCAGAA GCA AATTTTCTACTT ACGACCTCACTTTCTGTTGCGTTCTCT CCCTTCCCCTCCT			
Rabbit	TGATTTCTCG GCAGCCAGGGAGGG CCCCACGAC-AAGCCATT CAAAATCCCAGAA GT GATTTTCTACTT ACGACCTCACTTTCTGTTG---CTCT TCCTTCCCCTCCA			



20 bp dynamic
shifting window
>80% ID

~1 Meg region, 5q31

- Combination of [database searches](#) with [comparative sequence analysis](#) reduces the number of predicted transcription factor binding sites by several orders of magnitude

	Coding	Noncoding
Human interval Transfac predictions for <i>GATA</i> sites	839	20654
Aligned with the same predicted site in the mouse seq.	450	2618
Aligned sites conserved at 80% / 24 bp dynamic window	303	731
Random DNA sequence of the same length	29280	

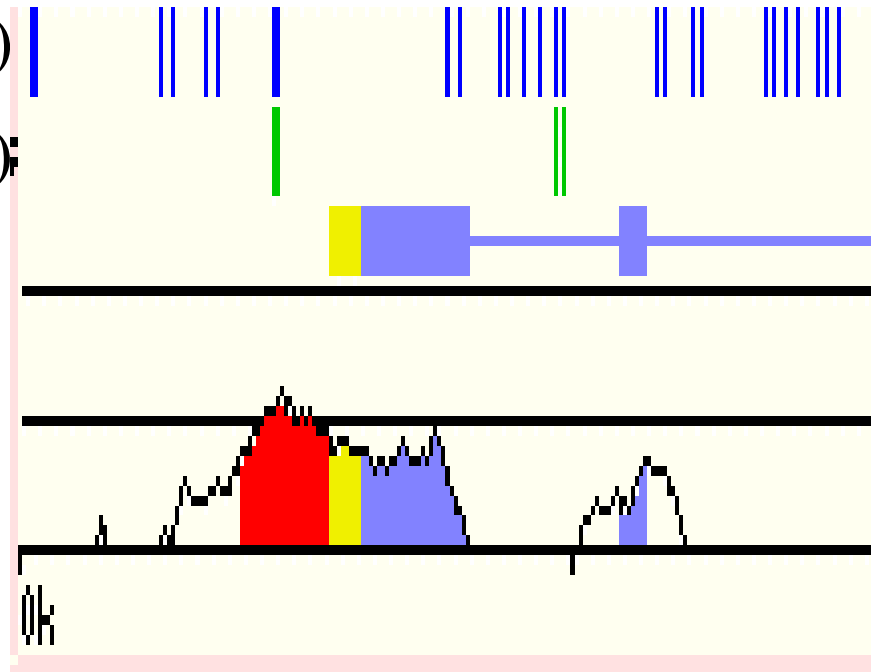
2 Exp. Verified GATA-3 Sites



IL 5

GATA-3 (28)

GATA-3 Conserved (4)



Choose Families to Visualize

<input type="checkbox"/> AHR	<input type="checkbox"/> E2	<input type="checkbox"/> LMO2COM	<input type="checkbox"/> RORA1
<input type="checkbox"/> AHRARNT	<input type="checkbox"/> E2F	<input type="checkbox"/> LYF1	<input type="checkbox"/> RORA2
<input type="checkbox"/> AML1	<input type="checkbox"/> E47	<input type="checkbox"/> MAX	<input type="checkbox"/> RREB1
<input checked="" type="checkbox"/> AP1	<input type="checkbox"/> E4BP4	<input type="checkbox"/> MEF2	<input type="checkbox"/> RSRFC4
<input type="checkbox"/> AP1FJ	<input type="checkbox"/> EGR1	<input type="checkbox"/> MIF1	<input type="checkbox"/> S8
<input type="checkbox"/> AP2	<input type="checkbox"/> EGR2	<input type="checkbox"/> MYB	<input type="checkbox"/> SEF1
<input checked="" type="checkbox"/> AP4	<input type="checkbox"/> EGR3	<input type="checkbox"/> MYCMAX	<input type="checkbox"/> SOX5
<input type="checkbox"/> ARNT	<input type="checkbox"/> ELK1	<input type="checkbox"/> MYOD	<input type="checkbox"/> SP1
<input type="checkbox"/> ARP1	<input type="checkbox"/> ER	<input type="checkbox"/> MZF1	<input type="checkbox"/> SREBP1
<input type="checkbox"/> ATF	<input type="checkbox"/> EVI1	<input type="checkbox"/> NF1	<input type="checkbox"/> SRF

Picture

Bases per layer:

Picture width (in pixels):

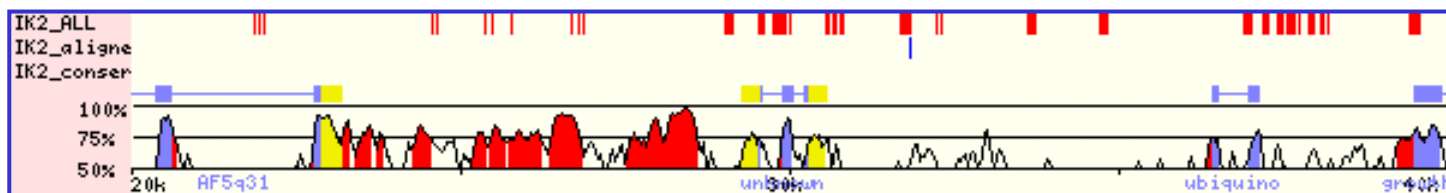
Clustering

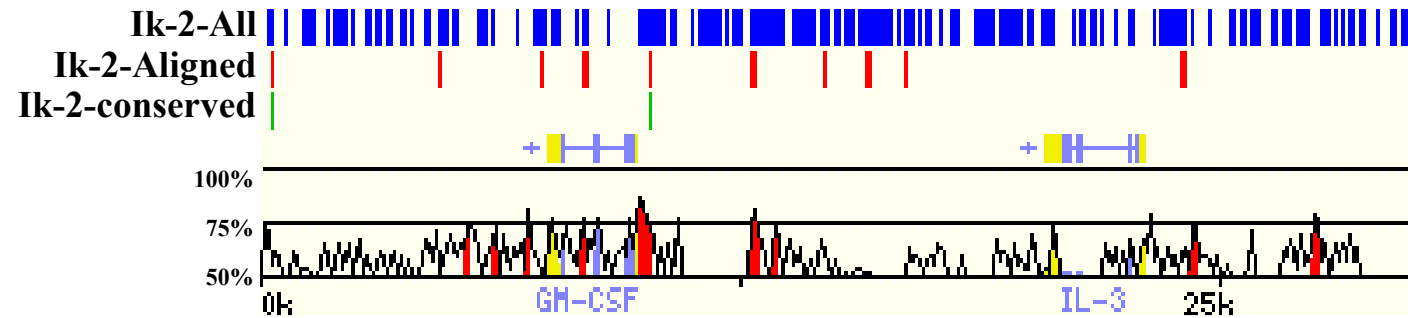
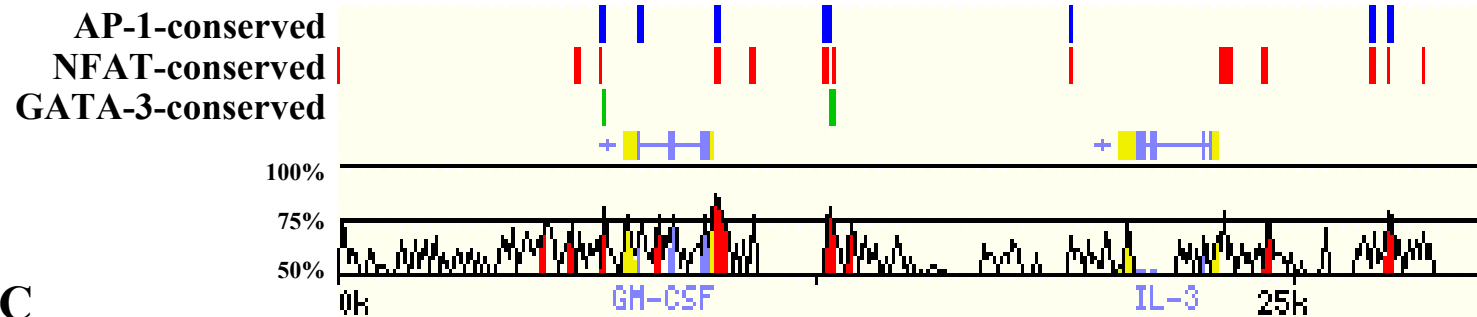
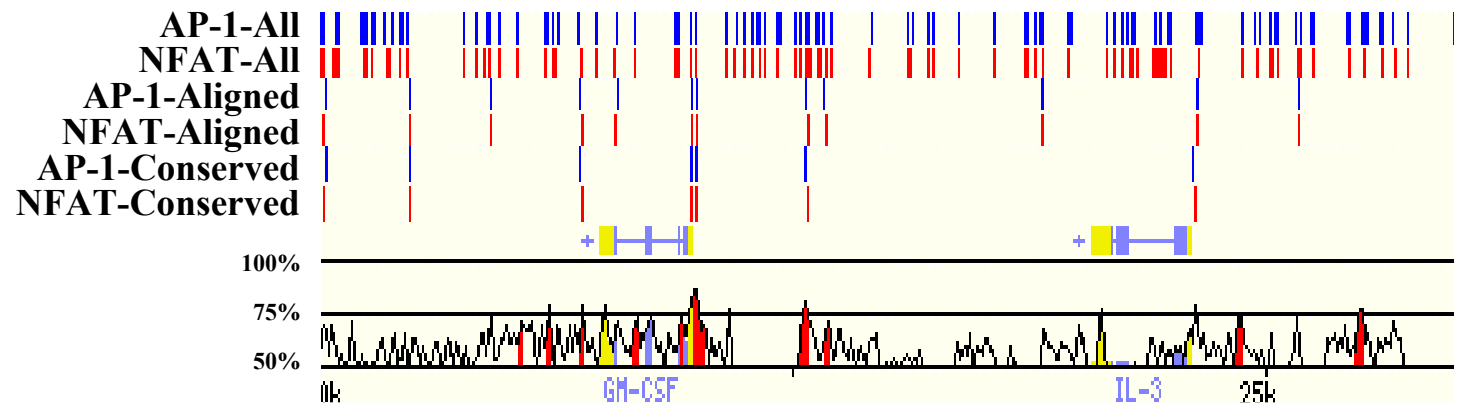
(Clustering: X sites over Y bases)
(1 site per N bases = NO clustering)

Types

☒ conserved
☐ aligned
☐ all

Select Picture Properties, Clustering, and Type of Visualization



A**B****C**

Sequence motif recognition

+

multiple sequence alignment of syntenic regions,



a high throughput strategy for filtering and prioritizing putative DNA binding sites



genomically informed starting place for globally investigating detailed regulation

Main features of VISTA

- Clear , configurable output
- Ability to visualize several global alignments on the same scale
- Alignments up to several megabases
- Working with finished and draft sequences
- Available source code and WEB site

Reviews on comparative genomics

- Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**: 369-72.
- Frazer, K.A, Elnitski, L., Church, D.M., Dubchak, I. , and Hardison, R.C.. Cross-species Sequence Comparisons: A Review of Methods and Available Resources. (2003) *Genome Res.*, 2003 Jan;13(1):1-12.
- Pennacchio LA, Rubin EM. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet*, 2001; 2:100-9.
- Wei, L., Liu, I., Dubchak, I. Shon, J., and Park, J. Comparative genomics approaches to study organism similarities and differences. *J Biomed Inform.*(2002) 35:142-50.

VISTA publications

- I. Dubchak, M. Brudno, L.S. Pachter, G.G. Loots, C. Mayor, E. M. Rubin, K. A. Frazer. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. *Genome Res.*, 10: 1304-1306.
- C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, Lior S. Pachter, I. Dubchak. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16: 1046-1047.
- Bray, N., Dubchak, I., and Pachter, L. AVID: A Global Alignment Program. (2003) *Genome Res.* 2003 Jan;13(1):97-102.
- G. G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak and E. M. Rubin. (2002) Comparative sequence-based approach to high-throughput discovery of functional regulatory elements. *Genome Res.*, 12:832-839

What if you don't have sequences of different species for the genomic region of your interest?

Are there publicly available comparative genomics data?

Large scale VISTA applications:

The Berkeley Genome Pipeline - comparing complete genomes

<http://pipeline.lbl.gov>

Cardiovascular comparative genomics database

<http://pga.lbl.gov>

THE BERKELEY GENOME PIPELINE

[FINISHED ANALYSIS](#) [ASSEMBLY ANALYSIS](#) [VISTA BROWSER](#) [VISTA TRACK](#) [MYGODZILLA SERVER](#) [SOFTWARE](#) [LINKS](#) [CONTACT INFO](#)

Automatic computational system for
comparative analysis of pairs of genomes

<http://pipeline.lbl.gov>

Alignments (all pair combinations):

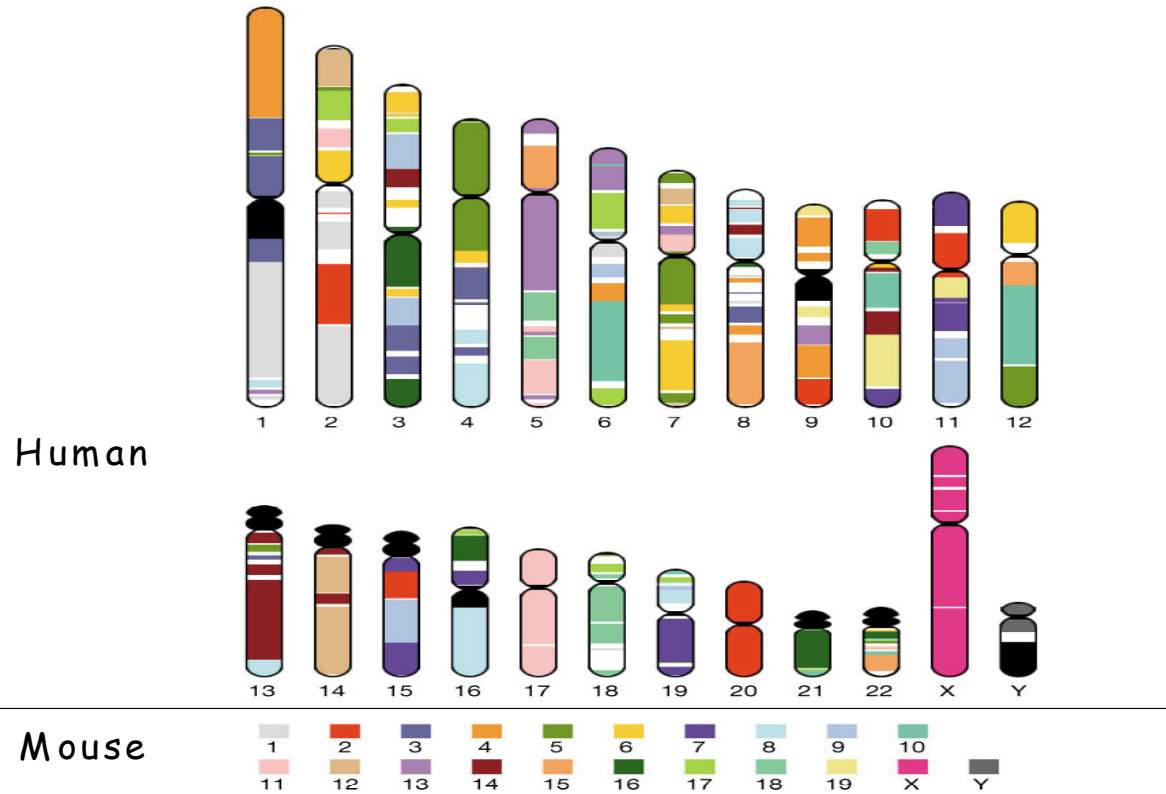
Human Genome: Golden Path Assembly)

Mouse assemblies: Arachne, Phusion (2001) MGSC v3 (2002)

Rat assemblies: November 2002, February 2003

D. Melanogaster vs D. Pseudoobscura February 2003

Chromosome Comparison



Base pair alignment

```

247  GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA  AGAGC
      |:  ||  ||||:  ||||  --:||  |||  |::|  |||---|||
368  GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG
  
```

Main modules of the system

Mapping and alignment of mouse contigs
against the human genome

```
graph TD; A[Mapping and alignment of mouse contigs against the human genome] --> B[Visualization]; A --> C[Analysis of conservation];
```

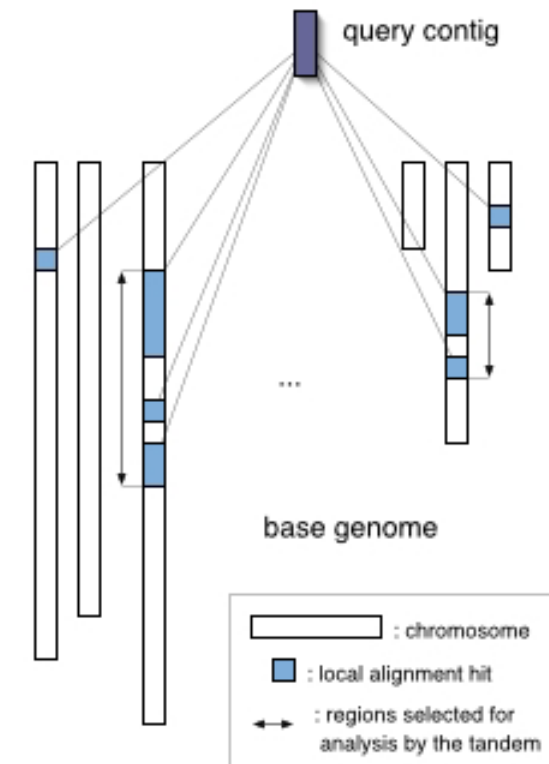
Visualization

Analysis of conservation

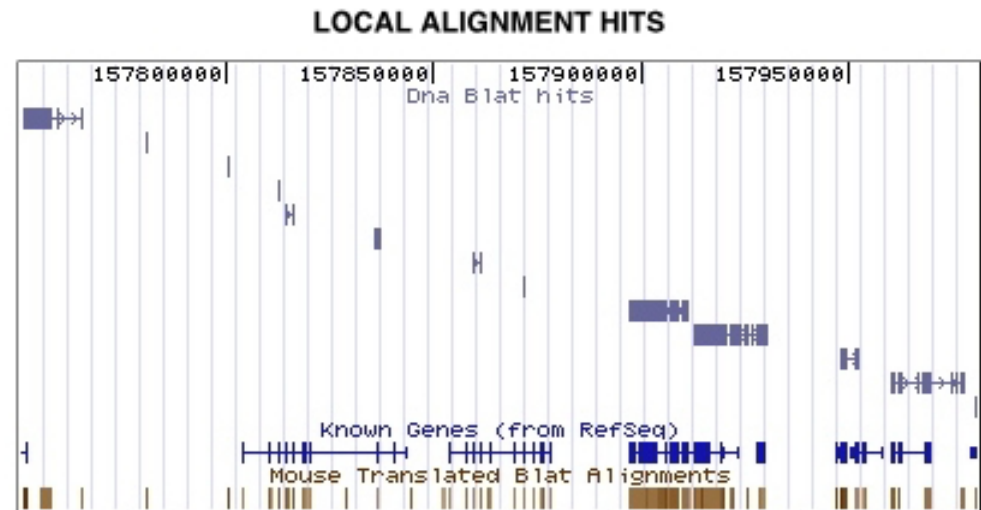
Tandem Local/Global Alignment Approach

Sequence fragment **anchoring** (DNA and/or translated BLAT)

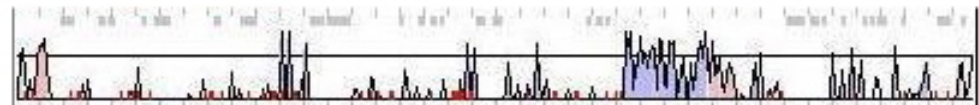
Multi-step verification of potential regions using global alignment (AVID or LAGAN)



ANCHORS FROM LOCAL ALIGNMENT HITS

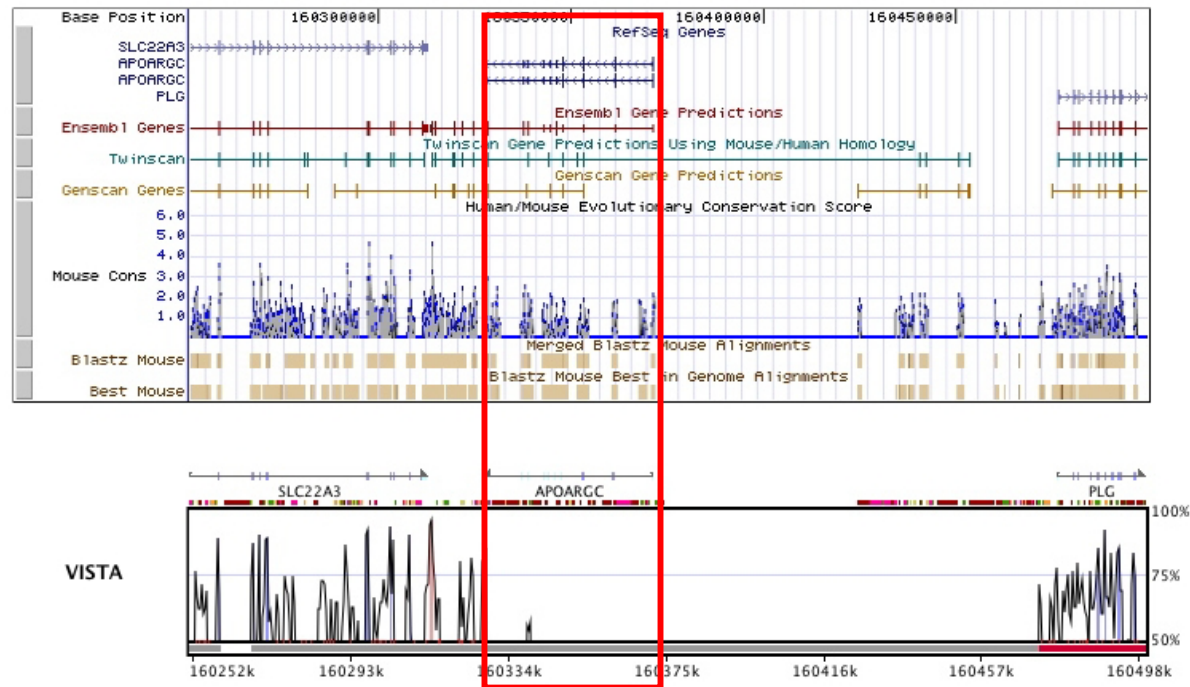


GLOBAL ALIGNMENT



Tandem approach in comparison with local alignment

Better specificity while preserving good sensitivity



Apolipoprotein(a) region. The expressed gene is confined to a subset of primates. Our method predicts that **apo(a)** has no homology in the mouse that local alignment can't detect.

VISTA Browser

Preprocessed whole genome comparison for pairs of species (human/mouse/rat & drosophilas)

THE BERKELEY GENOME PIPELINE

Compare the Human and Mouse Genomes

Please enter a gene name or a position (e.g. chrX:1-100000) on the Human Genome and press the "Go" button:

Base genome	Dataset	Position	
Human Nov. 2002	MGSCv3	ABCA1	Go

Description: Nov. 2002 Human Genome Assembly, NCBI build 31 (UCSC: hg13) vs Mouse Genome Sequencing Consortium, MGSCv3 (UCSC: mm2)

<http://pipeline.lbl.gov/>

VistaBrowser



Text browser

[HOME](#) [VISTA BROWSER](#) [GENOMEVISTA SERVER](#) [SOFTWARE](#) [CONTACT INFO](#)

Now Browsing

mouse Mouse Feb. 2002

human Human Nov. 2002

aligned with AVID

[<<](#) [>>](#)

Hits on chr9:99286788-99433941

[RefSeq in this region](#) [View in Vista Browser](#) [View at UCSC](#) [Get conserved regions](#)

mouse Contig info	Location on human	Alignment
chop250k 2219 Mapping = chr4(+):51737897-52144566 Contig Sequence (softmasked) length = 406670bp aligned: between 2719-400714 (397996bp)	chr9:99049271-99572723 Sequence (softmasked) RefSeq Conserved Regions length=523453bp	alignment

Text Browser

Select Genome Pair:

Mouse Feb. 2002 - Human Nov. 2002 (AVID) ▼

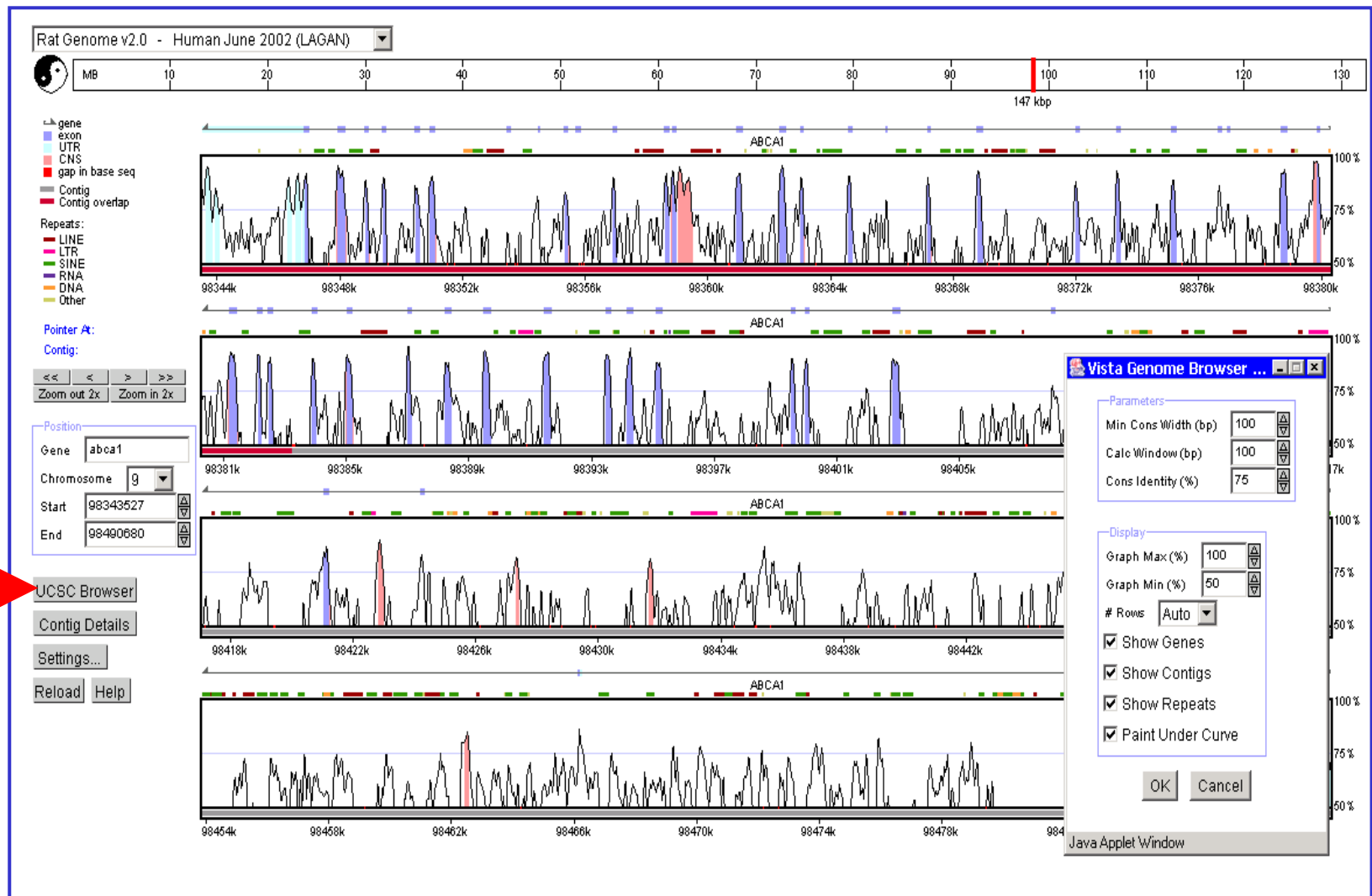
Position in the Base Genome:

(Format: chr11:113030619-113173035)

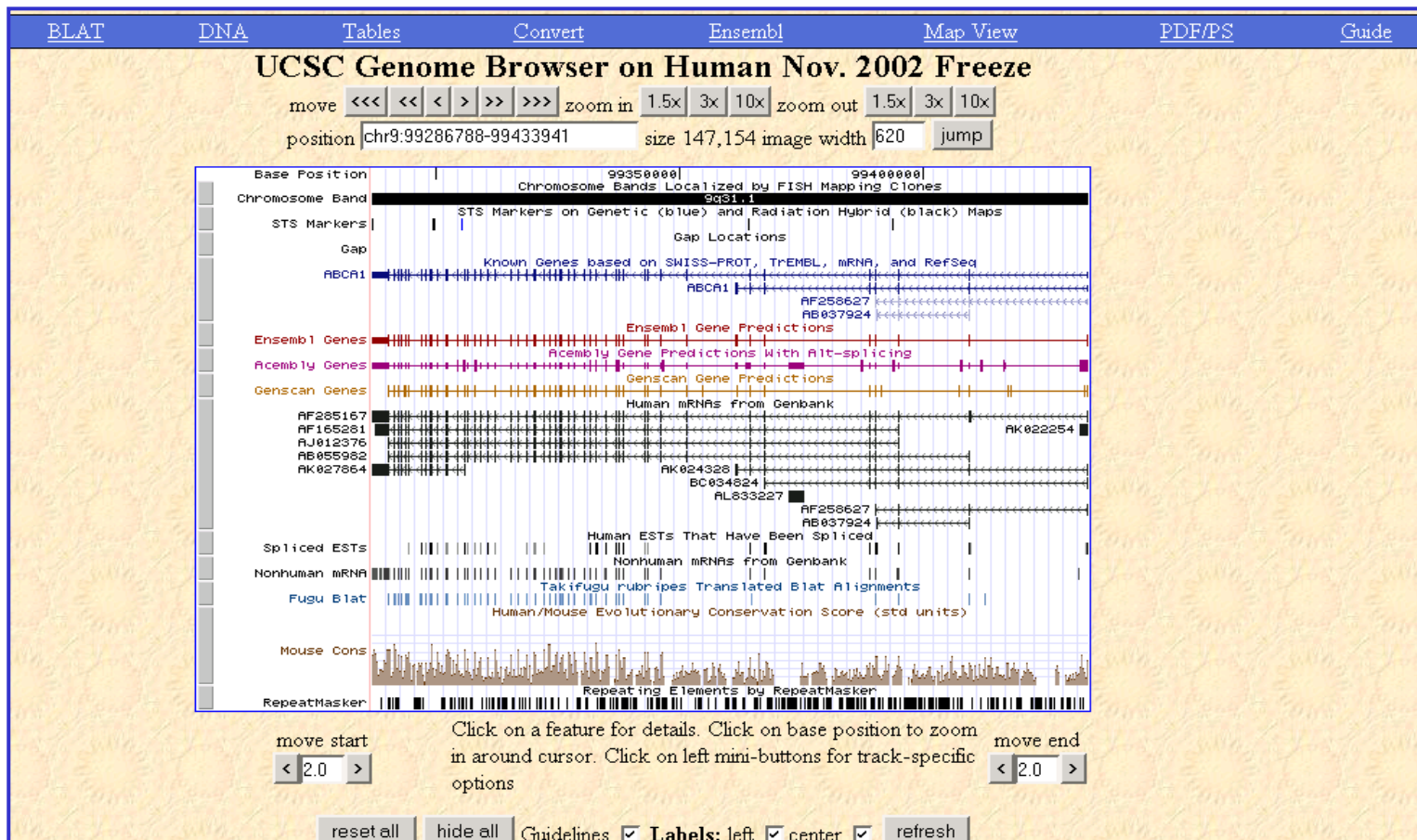
chr9:99286788-99433941

Go

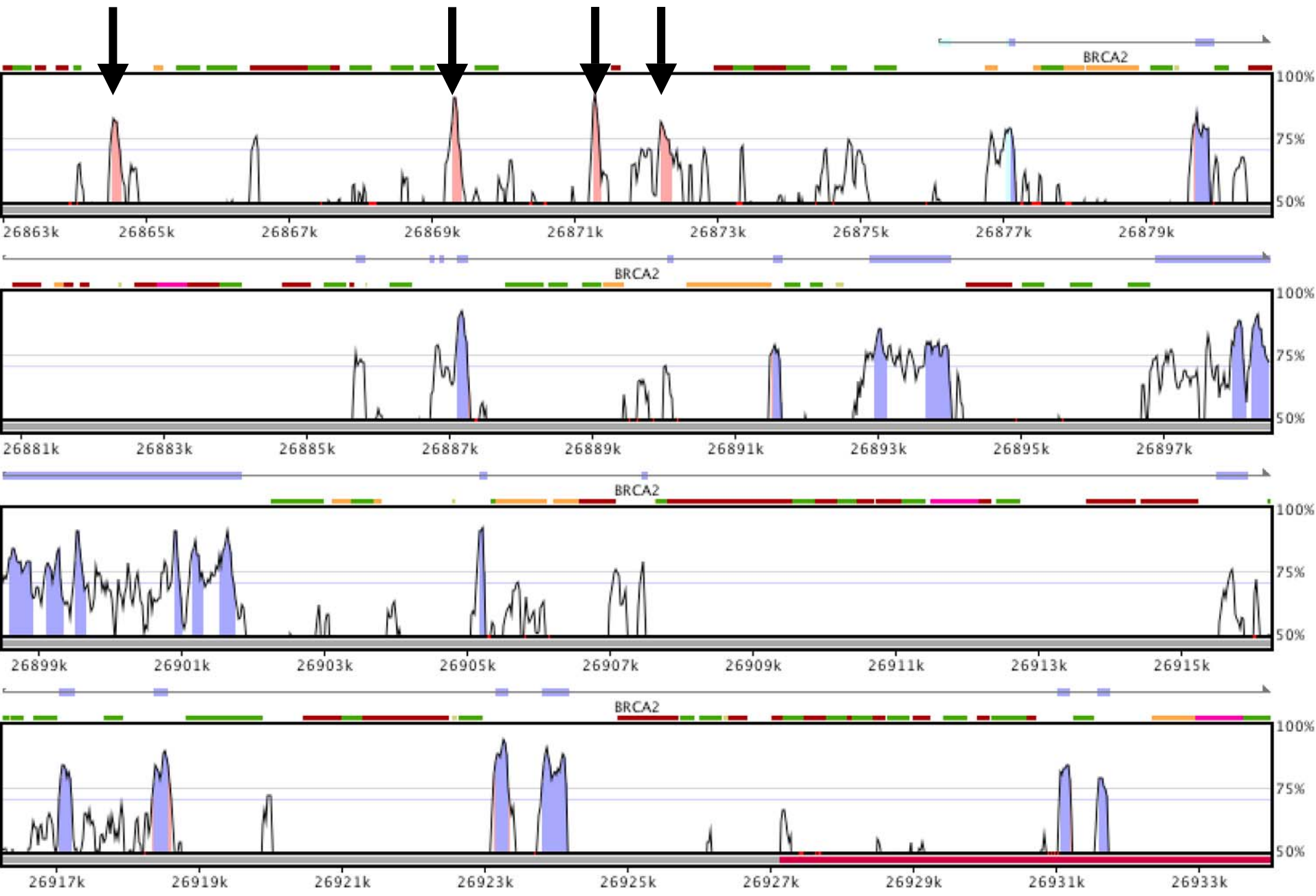
VistaBrowser



ABCA1 interval in UCSC human genome browser



VISTA Browser (Human/Mouse BRCA2 Comparison)



GenomeVista - is an interactive for comparing your favorite sequence against the base genome

THE BERKELEY GENOME PIPELINE

Compare the Human and Mouse Genomes

Please enter a gene name or a position (e.g. chrX:1-100000) on the Human Genome and press the "Go" button:

Base genome	Dataset	Position
Human Nov. 2002	MGSCv3	ABCA1

Go

Description: Nov. 2002 Human Genome Assembly, NCBI build 31 (UCSC: hg13) vs Mouse Genome Sequencing Consortium, MGSCv3 (UCSC: mm2)

<http://pipeline.lbl.gov/>

GenomeVISTA

Self-Input Sequence Comparison to either Human, Mouse, Rat, D.Melanogaster Reference Genomes

[HOME](#) [VISTA BROWSER](#) [GENOMEVISTA SERVER](#) [SOFTWARE](#) [CONTACT INFO](#)

GenomeVista

Submit a Request

Sequence
(choose one of the three options)

Paste a Query Sequence: (Paste a [Finished sequence](#) or [Draft sequence](#) in Fasta format , 300K max)

Alternatively, you can also select a file or enter a GenBank identification number:

FASTA

GenBank

Text files only. Word documents are **not** accepted. Sequences should be in FASTA format

☐ Treat lower-case letters as repeats

Base Genome

Advanced Options

Your E-mail
(we will inform you via e-mail when the results are available)

Name of request
(just something for you to identify the data set)

Organism

Human Nov. 2002
Rat Nov. 2002
D. melanogaster r.3
Mouse Feb. 2003
Human Dec. 2001
Mouse Feb. 2002
Human June 2002
Human Nov. 2002

default

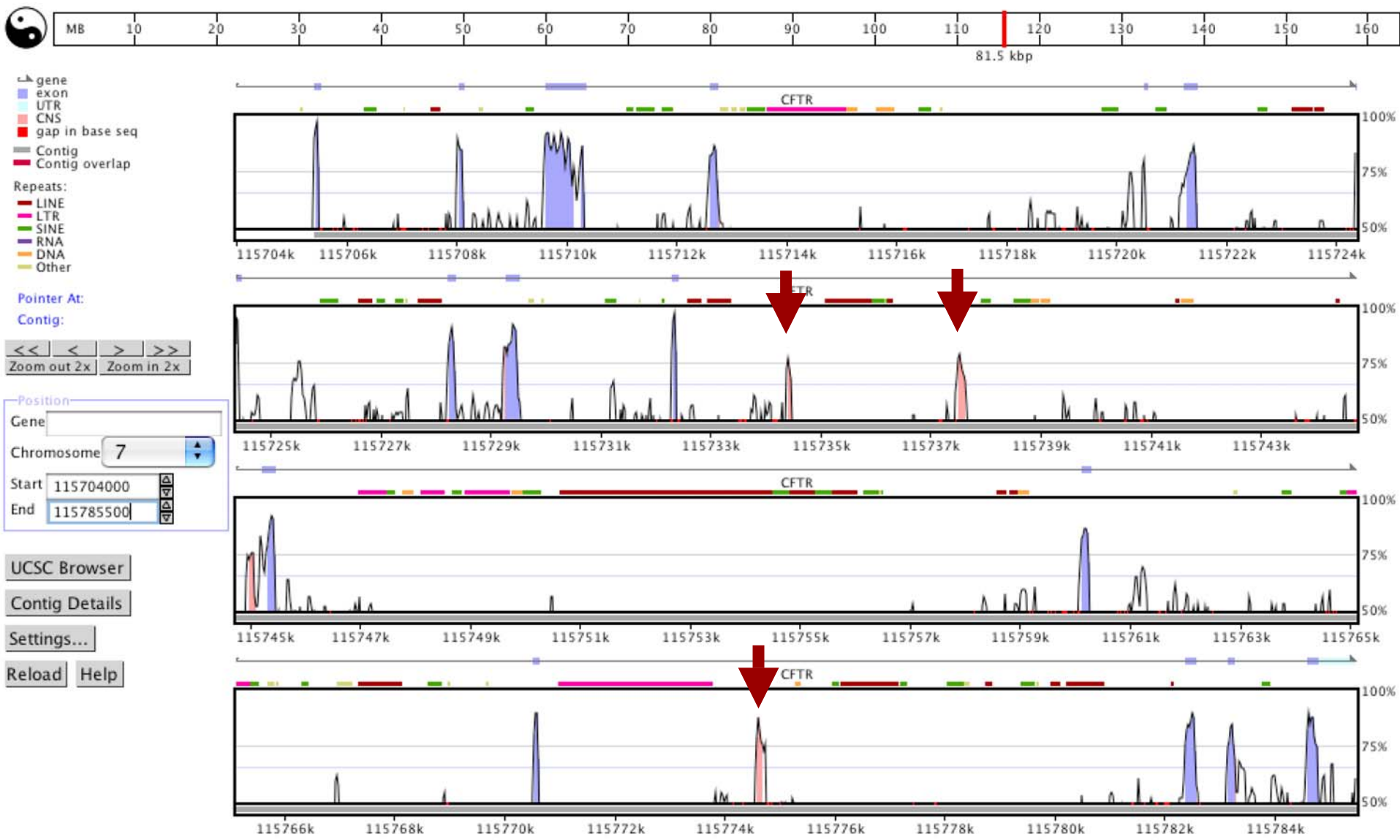


<http://pipeline.lbl.gov/>



GenomeVISTA

Random Opposum BAC versus Human Genome



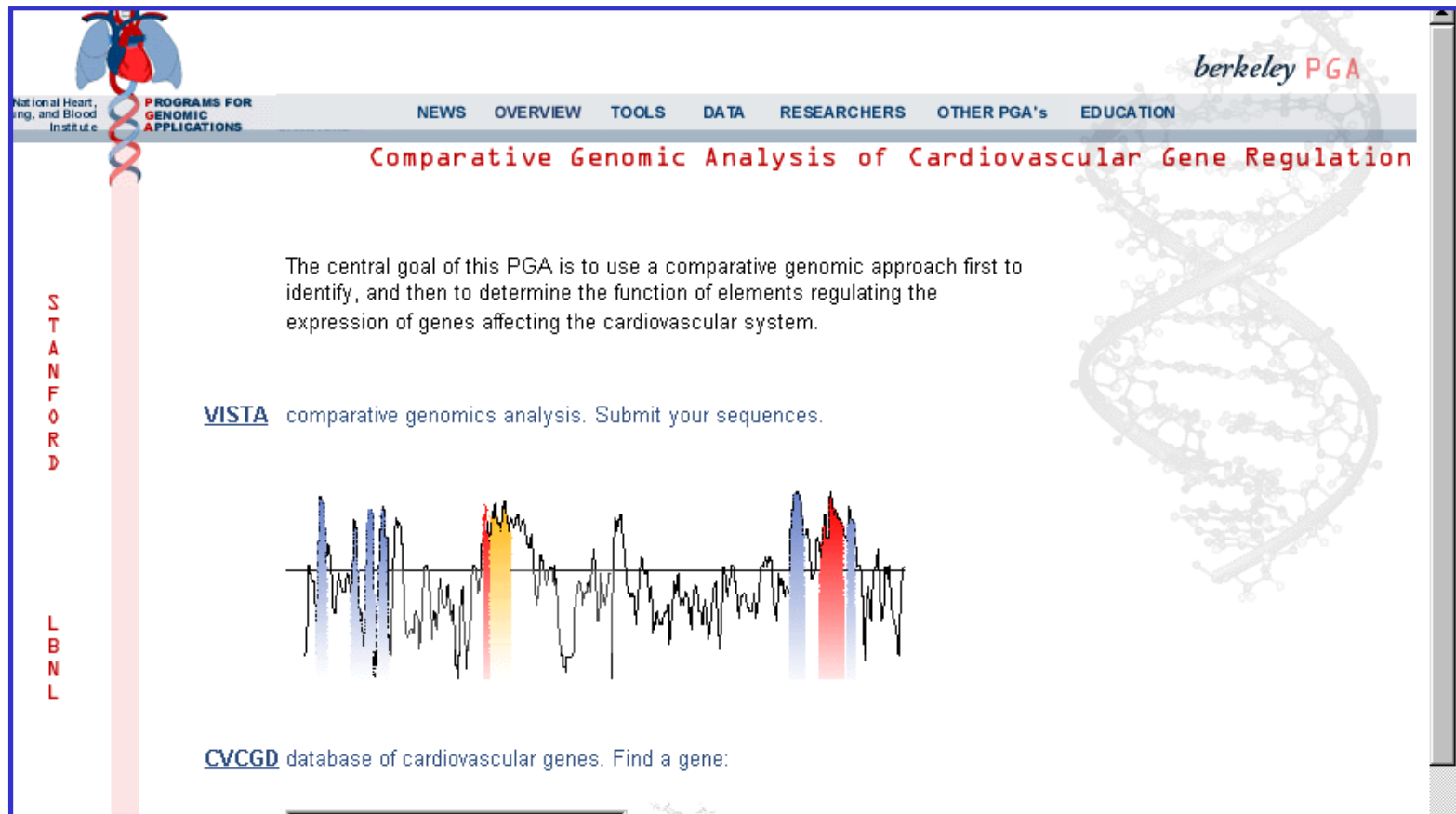
Results of an on-line submission of a draft unannotated platypus sequence AC130185 to Genome Vista. The gene has been correctly identified.



user query contig	location on human chr:start-end DNA = get DNA sequence RM = repeats masked RefSeq covered (if any) Conserved = 70% cons over 100bp	matches number of matches	
AC130185 platypus			
AC130185-4	chr7:117056599-117076707		
DNA	DNA RM RefSeq	5505	alignment
sequence total length = 35423bp	Conserved Regions		Vista
aligned: between 5997-24063 (18067bp)	length=20109bp		
AC130185-5	chr7:117095537-117127263		
DNA	DNA RM RefSeq	8980	alignment
sequence total length = 37422bp	Conserved Regions		Vista
aligned: between 5100-26725 (21626bp)	length=31727bp		

Comparative analysis of genomic intervals containing important cardiovascular genes

<http://pga.lbl.gov>



The screenshot shows the Berkeley PGA website. At the top left is the logo of the National Heart, Lung, and Blood Institute, featuring a heart and lungs. To its right is the text "PROGRAMS FOR GENOMIC APPLICATIONS". A navigation bar contains links: NEWS, OVERVIEW, TOOLS, DATA, RESEARCHERS, OTHER PGA's, and EDUCATION. The main title "Comparative Genomic Analysis of Cardiovascular Gene Regulation" is displayed in red. Below this, a paragraph states: "The central goal of this PGA is to use a comparative genomic approach first to identify, and then to determine the function of elements regulating the expression of genes affecting the cardiovascular system." Further down, there is a link to "VISTA" with the text "comparative genomics analysis. Submit your sequences." Below this is a line graph with several peaks, some of which are highlighted with colored vertical bars (blue, yellow, red). At the bottom, there is a link to "CVCGD" with the text "database of cardiovascular genes. Find a gene:". On the left side of the page, the words "STANFORD" and "LBL" are written vertically. On the right side, the words "berkeley PGA" are written in a stylized font. A large, faint DNA double helix is visible in the background on the right side.

National Heart, Lung, and Blood Institute

PROGRAMS FOR GENOMIC APPLICATIONS

NEWS OVERVIEW TOOLS DATA RESEARCHERS OTHER PGA's EDUCATION

berkeley PGA

Comparative Genomic Analysis of Cardiovascular Gene Regulation

The central goal of this PGA is to use a comparative genomic approach first to identify, and then to determine the function of elements regulating the expression of genes affecting the cardiovascular system.

VISTA comparative genomics analysis. Submit your sequences.

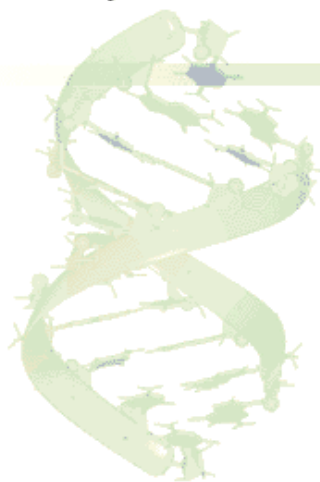
CVCGD database of cardiovascular genes. Find a gene:

STANFORD

LBL

<http://pga.lbl.gov/cvcgd.html>

berkeley PGA



Cardiovascular Comparative Genomic Database (CVCGD)

This database includes well-studied CV genes, for which an understanding of regulation should provide insights into CV relevant biological issues. While only a fraction of these genes will be characterized in the PGA biological projects over the 4-year time period of this program, the sequence of ~200 genomic intervals containing CV genes will be obtained and comparatively annotated and included in the CVCGD.

The database contains a variety of information for each gene relevant to this project:

- Gene name;
- Gene ID in the OMIM database (**OMIM**);
- Human map location (**HM**);
- GenBank accession number for human cDNA (**HC**);
- Mouse map location (**MM**);
- GenBank accession number for mouse cDNA (**MC**).

SEARCH the CVCGD

- [by gene name and abbreviation](#)
- [sorted alphabetically](#)
- [by categories](#) (groups of diseases).

Search Results

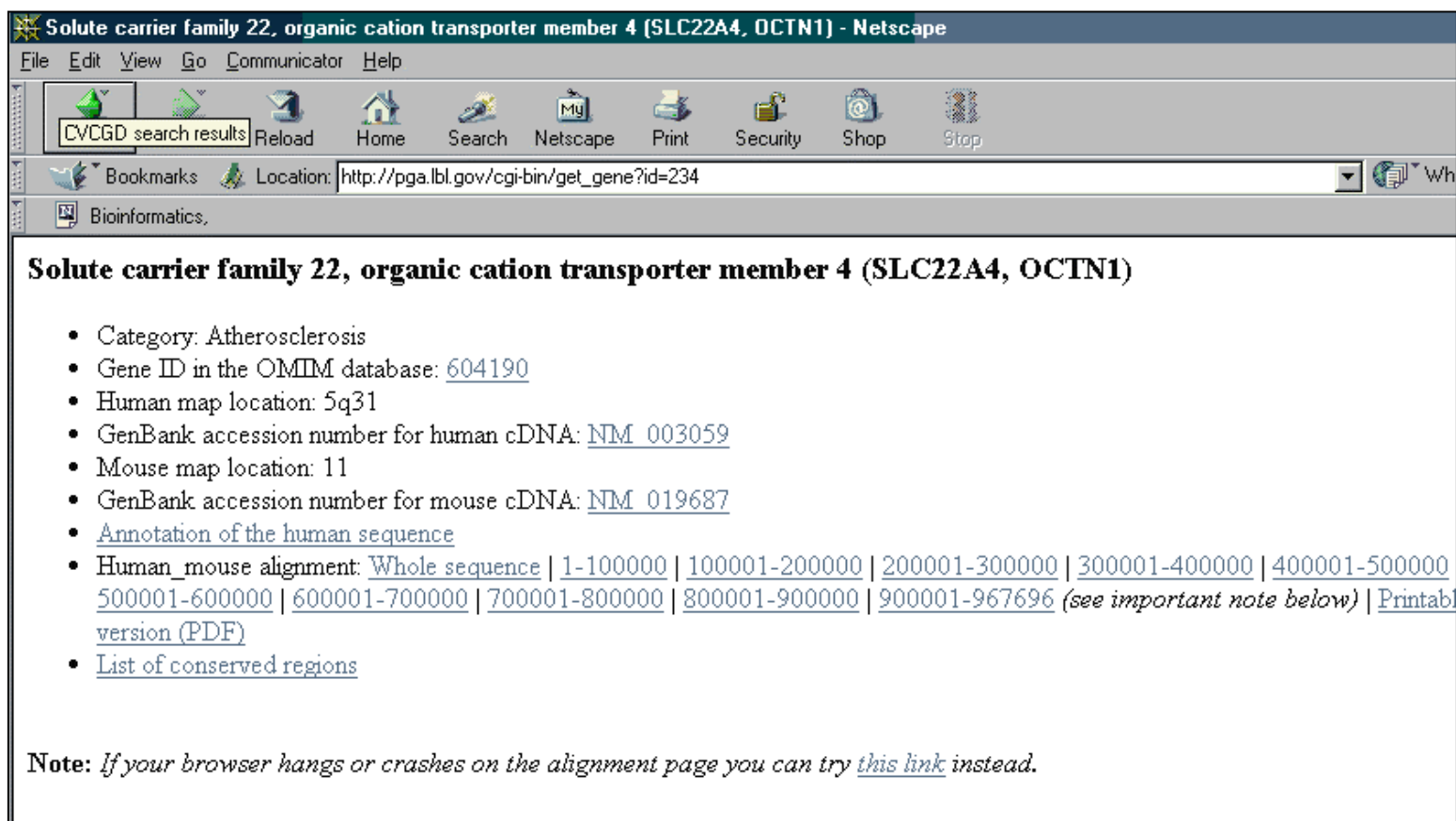
Links to whole
genome alignment

Table 1. Cardiovascular genes

Gene Name	Abbreviation	OMIM	HM	HC	MM	M
11-beta-hydroxysteroid dehydrogenase, type I	HSD11B1	600713	1p13.1	NM_005525		NM_008288
11-beta-hydroxysteroid dehydrogenase, type II	HSD11B2	218030	16q22	NM_000196		NM_008289
Acetyl-CoA acetyltransferase 1	ACAT1	203750	11q22.3-q23.1	NM_000019		
Acetyl-CoA acetyltransferase 2	ACAT2	100678	6q25.3-q26	NM_005891	17	M35797
Adducin 1	ADD1	102680	4p16.3	NM_001119	5	AF096839
Adducin 2	ADD2	102681	2p13-p14	X58199	6	AF100422
Adenosine A2 receptor	ADORA2A	102776	22q11.23	NM_000675		U05672
Adrenomedullin	ADM	103275	11p15.4	NM_001124	7	NM_009627
Agouti	ASIP					
Aldehyde reductase 1	AKR1B1 , ALDR1	103880	7q35	J04794		AF225564
Aldosterone synthase	CYP11B2	124080	8q21	NM_000498	15	NM_009991
Alpha myosin heavy chain	MYH6 , MYHCA	160710	14q12	NM_000257	14	M12290
Alpha tropomyosin	TPM1 , TMSA	191010	15q22.1	NM_000366	9	NM_009416
Alpha-1C-adrenergic receptor	ADRA1C	104221	8p21	NM_000680		AF031431
Angiopietin-1	ANGPT1	601667	8q22	NM_001146	15	U83509
Angiopietin-2	ANGPT2	601922	8q21	NM_001147	8	NM_007426
Angiotensin I converting enzyme/ kininase II	ACE , DCP1	106180	17q23	NM_000789	11	M55333
Angiotensin receptor 1	AGTR1	106165	3q21-q25	NM_000685		

Sequenced in Berkeley PGA

Example of CVCGD interval sequenced in Berkeley PGA

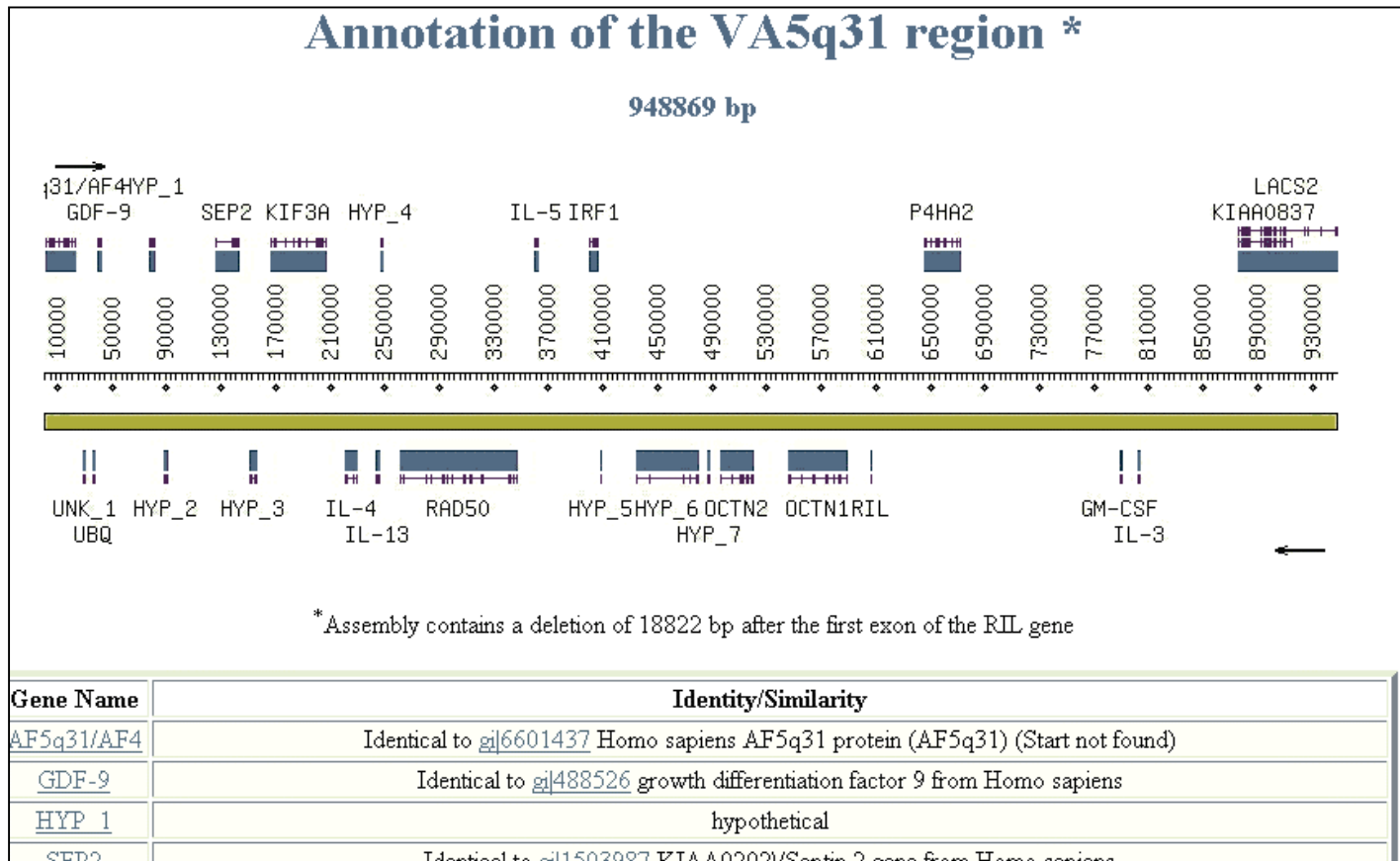


The screenshot shows a Netscape browser window with the title "Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1) - Netscape". The address bar displays the URL "http://pga.lbl.gov/cgi-bin/get_gene?id=234". The main content area is titled "Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)" and contains a list of gene-related information:

- Category: Atherosclerosis
- Gene ID in the OMIM database: [604190](#)
- Human map location: 5q31
- GenBank accession number for human cDNA: [NM_003059](#)
- Mouse map location: 11
- GenBank accession number for mouse cDNA: [NM_019687](#)
- [Annotation of the human sequence](#)
- Human_mouse alignment: [Whole sequence](#) | [1-100000](#) | [100001-200000](#) | [200001-300000](#) | [300001-400000](#) | [400001-500000](#) | [500001-600000](#) | [600001-700000](#) | [700001-800000](#) | [800001-900000](#) | [900001-967696](#) (see important note below) | [Printable version \(PDF\)](#)
- [List of conserved regions](#)

Note: *If your browser hangs or crashes on the alignment page you can try [this link](#) instead.*

Short annotation of the region



Summary

- VISTA family of tools
<http://www-gsd.lbl.gov/vista>
- PhyloVISTA
<http://www-gsd.lbl.gov/phylovista>
- Precomputed whole-genome alignments
<http://pipeline.lbl.gov>
- Berkeley PGA <http://pga.lbl.gov>

We'll be happy to work with you on your data
email - [ildubchak @lbl.gov](mailto:ildubchak@lbl.gov)

Publications on whole genome alignments:

- I.Dubchak, L. Pachter. (2002) The computational challenges of applying comparative-based computational methods to whole genomes. *Briefings in Bioinformatics*, 3, 18.
- Couronne O., Poliakov A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter L, Dubchak, I. (2002) Strategies and Tools for Whole Genome Alignments, *Genome Res.*, 2003 Jan;13(1):73-80.
- Waterston, et.al., Initial sequencing and comparative analysis of the mouse genome. *Nature*. (2002) 420:520-62.

Related sites

- The UCSC Genome Browser & BLAT program
<http://genome.ucsc.edu/>
- ENSEMBLE Project (Sanger Center) <http://www.ensembl.org/>
- AVID alignment program
<http://baboon.math.berkeley.edu/~syntenic/avid.html>
- SLAM comparative gene prediction program
<http://bio.math.berkeley.edu/slam/mouse/>
- PSU group's MHC Human-Mouse comparison results
<http://bio.cse.psu.edu/mousegroup/MHC/>
- PSU Pipmaker program <http://bio.cse.psu.edu/pipmaker/>

Towards Better VISTAs

Information
from a Single
Sequence
Alone



Multi-Organism
High Quality
Sequences



Thanks

Biology

Kelly Frazer
Gaby Loots
Len Pennacchio

Eddy Rubin

Bioinformatics

Michael Brudno
Olivier Couronne
Brian Klock
Chris Mayor
Ivan Ovcharenko
Alexander Poliakov
Jody Schwartz
Lior Pachter (UCB)

Funding - Programs for Genomic Applications (PGA) by NHLBI